

**PROCEEDINGS IN
INFORMATICS AND INFORMATION TECHNOLOGIES**

**11TH WORKSHOP ON
INTELLIGENT AND KNOWLEDGE
ORIENTED TECHNOLOGIES
35TH CONFERENCE ON
DATA AND KNOWLEDGE**

MÁRIA BIELIKOVÁ, IVAN SRBA (Eds.)

Proceedings in
Informatics and Information Technologies

WIKT & DaZ 2016
11th Workshop on Intelligent
and Knowledge Oriented Technologies
35th Conference on Data and Knowledge

Mária Bielíková and Ivan Srba (Eds.)

WIKT & DaZ 2016

11th Workshop on Intelligent
and Knowledge Oriented Technologies
35th Conference on Data and Knowledge

Proceedings

November 3-4, 2016
Smolenice, Slovakia

Proceedings in Informatics and Information Technologies
WIKT & DaZ 2016

Editors

Mária Bielíková, Ivan Srba

Faculty of Informatics and Information Technologies,
Slovak University of Technology in Bratislava
Ilkovičkova 2, 842 16 Bratislava, Slovakia
{maria.bielikova, ivan.srba}@stuba.sk

WIKT & DaZ 2016 was supported by

- project APVV-15-0508, HIBER – Human Information Behavior in the Digital Space
- project VEGA 1/0752/14, Intelligent analysis of big data by semantic-oriented and bio-inspired methods in a parallel environment
- project VEGA 1/0646/15, Adaptation of access to information and knowledge artifacts based on interaction and collaboration within web environment
- the Czech Society for Cybernetics and Informatics (CSKI)

© 2016 The authors listed in the Table of Contents

All contributions were reviewed by the Programme Committee and printed as delivered by authors without substantial modifications

Visit WIKT & DaZ 2016 on the Web: <https://wikt-daz2016.fiit.stuba.sk>

Executive Editor: Ivan Srba

Copy Editors: Ondrej Kaššák

Published by

Nakladateľstvo STU

Vazovova 5, Bratislava, Slovakia

ISBN 978-80-227-4619-9

Preface

Intelligent and knowledge-oriented technologies currently affect various areas of human lives. They form an important component of research activity of several research groups active in Slovakia and Czech Republic. They also constituted a subject of presentations and discussions in the Smolenice Castle, where the 11th Workshop on Intelligent and Knowledge oriented Technologies (WIKT) and 35th Conference on Data and Knowledge (Data a znalosti, DaZ) was held from the 3rd to the 4th of November 2016.

This year followed the tradition started in 2006 (the first WIKT workshop) and in 1981 (the first DATASEM conference, which precedes Data a znalosti). A series of workshops and conferences during the last years fostered the creative environment and research by making a forum for exchanging knowledge and creative discussions in the field of intelligent and knowledge oriented technologies in Slovakia and Czech Republic. The aim of WIKT & DaZ was always to bring together researchers from several research centres in Slovakia, Czech Republic and vicinity.

Main topics of WIKT workshop were:

- knowledge technologies and their applications,
- information and knowledge modeling, semantic representation,
- analysis and processing of information sources,
- social web and its applications, analysis of social networks,
- personalized web and its applications, recommendation,
- processing of information sources in Slovak language,
- semantic and service oriented architecture,
- reasoning and inference.

Main topics of Data a znalosti conference were:

- data mining,
- machine learning, classification and prediction systems,
- creation, publication and employment of open data,
- indexing and retrieval text and multimedia data,
- user modelling, adaptive and personalized systems,
- advanced user interfaces of software and information systems,
- systems for knowledge management in organizations,
- expert, intelligent and agent systems,
- natural language processed applied at real tasks,
- ontologies and conceptual models applied at real tasks,
- automatic reasoning and planning applied at real tasks.

Authors sent their contributions in the form of extended abstracts (in Slovak, Czech and English) of the following types:

- research paper,
- work-in-progress paper,
- application paper,
- position paper,
- PhD symposium (a special challenge for doctoral students who could offer a contribution related to the direction and goals of their dissertation).

The workshop WIKT and Data a znalosti conference reaffirmed its significance this year, again. A total of 52 papers were submitted, most of them as research paper or work-in-progress paper. Each contribution was reviewed by three members of the program committee. The result of the assessment was acceptance of 50 papers in total. All papers were presented in lively style of short presentations followed by poster discussions. 10 of them were accepted for longer presentation, 26 for short announcement. 14 submission were accepted for PhD symposium.

Following DaZ conference tradition 8 invited lectures (among them four experts from industry) on interesting topics of information processing were presented.

Majority of the papers are written in the native language of the authors, i.e., in Slovak or Czech. The language of the workshop was Slovak and Czech. This fact on the one hand limits the dissemination of the results, but on the other hand it helps in growing professional language skills in the domain of rapidly developing information, knowledge and web technologies.

We continued also with good tradice to organize a project meeting just before the conference. The meeting of HIBER project (Human Information Behavior in the Digital Space) was held on November 2nd, 2016. 32 researchers from Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava and Faculty of Arts, Comenius University discussed and brainstormed in groups on the beginning project directions.

We are very pleased that this year Smolenice Castle was a record in number of participating research groups from Slovakia and the Czech Republic. We thank all authors for interesting contributions initiating fruitful debates. We thank the members of the program committee, who willingly participated in the judging of submissions and discussions about the direction of the workshop. We also thank them for the contribution to the maintenance of high professional level of the event and the fact that they came to the workshop with their research groups.

We thank especially *Ondrej Kaššák* for preparing this proceedings and to all the members of the organizing committee, who made a considerable effort to turn a picturesque spot in the heart of Central Europe into a two day passionate scientific debate centre and helped to spread the knowledge and collaboration.

Bratislava, October 2016

Mária Bieliková and Ivan Srba

Predhovor

Inteligentné a znalostne orientované technológie ovplyvňujú v súčasnosti najrôznejšie oblasti ľudskej činnosti. Tvorí aj významnú zložku náplne činnosti viacerých výskumných skupín pôsobiach na Slovensku a v Česku. Tvorili aj hlavnú tému prezentácií a diskusií na Smolenickom zámku 3. – 4. novembra, 2016, kde sa konal 11. ročník tvorivej pracovnej dielne o inteligentných a znalostne orientovaných technológiách WIKT 2016 v spojení s 35. ročníkom konferencie Dat a znalosti.

Tento ročník nadviazal na tradíciu započatú v roku 2006 (prvý ročník tvorivej dielne WIKT) a v roku 1981 (prvý ročník konferencie DATASEM, ktorá predchádzala konferencie Data a znalosti). Séria pracovných dielní a konferencií počas posledných rokov vytvorila tvorivé prostredie pre podporu výskumu najmä prostredníctvom výmeny poznatkov a tvorivých diskusií v atraktívnych oblastiach inteligentných a znalostne orientovaných technológií na Slovensku. Snahou dielne WIKT a konferencie Data a znalosti vždy bolo spájať výskumníkov viacerých výskumných centier v širšom zábere Slovenska a Českej republiky.

Hlavné témy tvorivej dielne WIKT 2016 boli:

- znalostné technológie a ich aplikácie,
- modelovanie informácií a znalostí, reprezentácia sémantiky,
- analýza a spracovanie informačných zdrojov,
- sociálny web a jeho aplikácie, analýza sociálnych sietí,
- personalizovaný web a jeho aplikácie, odporúčanie,
- spracovanie informačných zdrojov v slovenskom jazyku,
- sémanticky a servisne orientované architektúry,
- usudzovanie a odvodzovanie.

Hlavné témy konferencie Data a znalosti 2016 boli:

- dolovanie v dátach,
- strojové učenie, klasifikačné a prediktívne systémy,
- tvorba, publikovanie a využívanie otvorených a prepojených dát,
- indexovanie a vyhľadávanie textových a multimediálnych dát,
- modelovanie používateľa, adaptívne a personalizované systémy,
- pokročilé používateľské rozhrania softvérových a informačných systémov,
- systémy pre správu znalostí v organizáciách,
- expertné, inteligentné a agentové systémy, výpočtová inteligencia,
- výpočtová lingvistika aplikovaná na reálne úlohy,
- ontologické a konceptuálne modely aplikované na reálne úlohy,
- automatické odvodzovanie a plánovanie aplikované na reálne úlohy.

Autori zasielali príspevky v tvare rozšíreného abstraktu v slovenskom, českom alebo anglickom jazyku nasledujúcich kategórií:

- výskumný príspevok,
- príspevok o prebiehajúcom výskume,
- aplikačný príspevok,
- vizionársky príspevok,
- doktorandské sympóziu (špeciálnu výzvu mali študenti tretieho stupňa okolo dizertačnej skúšky, ktorí mohli ponúknuť príspevok o smerovaní a cieľoch svojej dizertačnej práce).

Tvorivá dielňa WIKT a konferencia Data a znalosti v tomto roku znovu potvrdila svoje opodstatnenie. Celkovo bolo ponúknutých 52 príspevkov, väčšina z nich v kategórii výskumný príspevok alebo príspevok o prebiehajúcom výskume. Každý príspevok posúdili traja členovia programového výboru. Výsledkom posudzovania bolo rozhodnutie o prijatí 50 príspevkov. Všetky príspevky autori prezentovali živou diskusiou pri posteroch, ktorej predchádzalo oznámenie o výsledkoch. 10 z nich bolo prijatých na dlhšie oznámenie, 26 na krátke oznámenie. 14 príspevkov bolo prijatých do doktorandskej sekcie, kde v štyroch sekciách prebehli zaujímavé diskusie.

Podľa tradícií konferencie Data a znalosti v programe bolo 8 pozvaných prednášok na zaujímavé témy spracovania informácií (medzi nimi boli štyria experti z priemyslu).

Väčšina príspevkov je napísaná v materinskom jazyku autorov, teda slovensky, resp. česky. Jazyk tvorivej dielne bol slovenský a český, čo síce na jednej strane ohraničuje šírenie výsledkov, ale na strane druhej pomáha pestovaniu odborného jazyka v doméne rýchlo rozvíjajúcich sa znalostných a aj webových technológií.

Pokračovali sme tiež v dobrej tradícii organizovania projektového stretnutia pred samotnou konferenciou. V stredu 2. novembra 2016 sa uskutočnilo stretnutie k projektu HIBER (Human Information Behavior in the Digital Space), na ktorom sa zúčastnili výskumníci z Fakulty informatiky a informačných technológií Slovenskej technickej univerzity v Bratislave a z Filozofickej fakulty Komenského univerzity. Ide o začínajúci projekt, takže hlavným cieľom bolo diskutovať o konkrétnom smerovaní projektu.

Záujem o tvorivú dielňu WIKT a konferenciu Data a Znalosti bol tento rok rekordným. Sme veľmi potešení, že na Smolenickom zámku bol rekordný počet výskumných skupín, konkrétne 12 pracovísk s viac ako jedným účastníkom. Ďakujeme všetkým autorom za zaujímavé príspevky podnecujúce diskusiu. Ďakujeme členom programového výboru, ktorí ochotne participovali na posudzovaní príspevkov a diskusiách o smerovaní tvorivej dielne. A tiež za príspevok k udržaniu vysokej odbornej úrovne celého podujatia aj tým, že na pracovnú dielňu prišli aj so svojimi výskumnými skupinami.

Zároveň ďakujeme *Ondrejovi Kaššákovi* za perfektnú prácu pri príprave tohto zborníka a tiež všetkým členom organizačného výboru, ktorí vynaložili nemalé úsilie na to, aby sa dva dni jedno malebné miestečko v srdci strednej Európy premenilo na zanietené vedecké diskusie a pomohlo tak v šírení poznatkov a spolupráci.

Bratislava, október 2016

Mária Bieliková a Ivan Srba

Conference Organization

The 11th Workshop on Intelligent and Knowledge Oriented Technologies (WIKT) and 35th Conference on Data and Knowledge (Data a Znalosti), held on November 3-4, 2016 in Smolenice, was organised by the Slovak University of Technology in Bratislava (and, in particular, its Institute of Informatics, Information Systems and Software Engineering, Faculty of Informatics and Information Technologies) in collaboration with Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University Košice and the Institute of Informatics, Slovak Academy of Sciences, Bratislava under considerable support of Informatics Development Foundation.

Programme Committee

Chair

Mária Bielíková, FIIT STU, Bratislava

Members

František Babič
FEI TU, Košice

Daniela Chudá
FIIT STU, Bratislava

Michal Barla
FIIT STU, Bratislava

Martin Dostál
ZČU, Plzeň

Roman Barták
MFF UK, Praha

Igor Farkaš
FMFI UK, Bratislava

Vladimír Bartík
FIT VUT, Brno

Dalibor Fiala
ZČU, Plzeň

Petr Berka
FIS VŠE, Praha

Ján Genčí
FEI TU, Košice

Přemek Brada
ZČU, Plzeň

Peter Gurský
PrF UPJŠ, Košice

Radek Burget
FIT VUT, Brno

Zdeněk Havlice
FEI TU, Košice

Peter Butka
FEI TU, Košice

Ladislav Hluchý
SAV, Bratislava

Dušan Chlapek
FIS VŠE, Praha

Martin Holeňa
Leibniz Institute, University of Rostock

Irena Holubová
MFF UK, Praha

Martin Homola
FMFI UK, Bratislava

Tomáš Horváth
PrF UPJŠ, Košice

Petr Hujňák
Per Partes Consulting, Praha

Jana Klečková
ZČU, Plzeň

Jiří Kléma
FEL ČVUT, Praha

Michal Kompan
FIIT STU, Bratislava

Pavel Kordík
FIT CVUT, Praha

Stanislav Krajčí
PrF UPJŠ, Košice

Pavel Král
ZČU, Plzeň

Petr Křemen
FEL ČVUT, Praha

Martin Labský
IBM Research, Praha

Peter Lacko
FIIT STU, Bratislava

Michal Laclavík
Magnetic, a.s.

Vitaly Levashenko
FRI ŽU, Žilina

Aleš Limpouch
TopoL Software

Marián Mach
FEI TU, Košice

Kristína Machová
FEI TU, Košice

Jan Martinovič
FEI VŠB, Ostrava

Karol Matiaško
FRI ŽU, Žilina

Peter Mikulecký
Univerzita Hradec Králové

Martin Molhanec
FEL ČVUT, Praha

Roman Moucek
ZČU, Plzeň

Pavol Návrát
FIIT STU, Bratislava

Martin Nečaský
MFF UK, Praha

Giang Nguyen
UI SAV, Bratislava

Marek Obitko
Rockwell Automation, Praha

Ján Paralič
FEI TU, Košice

Robert Pergl
FIT ČVUT, Praha

Tomáš Pitner
FI MU, Brno

Jaroslav Pokorný
MFF UK, Praha

Lubomír Popelínský
FI MU, Brno

Jaroslav Porubán
FEI TU, Košice

Jan Rauch
FIS VŠE, Praha

Václav Řepa
FIS VŠE, Praha

Karel Richta
FEL ČVUT, Praha

Viera Rozinajová
FIIT STU, Bratislava

Hana Rudová
FI MU, Brno

Petr Šaloun
FEI VŠB, Ostrava

Pavel Smrž
FIT VUT, Brno

Václav Snášel
FEI VŠB, Ostrava

Ivan Srba
FIIT STU, Bratislava

Josef Steinberger
ZČU, Plzeň

Július Štuller
Ústav informatiky AV ČR

Vojtěch Svátek
FIS VŠE, Praha

Jozef Tvarožek
FIIT STU, Bratislava

Michal Valenta
FIT ČVUT, Praha

Peter Vojtáš
MFF UK, Praha

Ondřej Zamazal
FIS VŠE, Praha

Jaroslav Zendulka
FIT VUT, Brno

Steering Committee WIKT

Mária Bielíková
FIIT STU, Bratislava

Peter Butka
FEI TU, Košice

Ladislav Hluchý
ÚI SAV, Bratislava

Martin Homola
FMFI UK, Bratislava

Tomáš Horváth
PrF UPJŠ, Košice

Daniela Chudá
FIIT STU, Bratislava

Stanislav Krajčí
PrF UPJŠ, Košice

Michal Laclavík
Magnetic, a.s.

Marián Mach
FEI TU, Košice

Viera Rozinajová
FIIT STU, Bratislava

Kristína Machová
FEI TU, Košice

Petr Šaloun
FEI VŠB-TU, Ostrava

Karol Matiaško
FRI ŽU, Žilina

Peter Vojtáš
MFF UK, Praha

Pavol Návrat
FIIT STU, Bratislava

Jaroslav Zendulka
FIT VUT, Brno

Ján Paralič
FEI TU, Košice

Steering Committee Data and Knowledge

Mária Bieliková
FIIT STU, Bratislava

Jaroslav Pokorný
MFF UK, Praha

Tomáš Horváth
PrF UPJŠ, Košice

Ľuboš Popelínský
FI MU, Brno

Petr Hujňák
Per Partes Consulting, Praha

Jan Rauch
FIS VŠE, Praha

Dušan Chlapek
FIS VŠE, Praha

Karel Richta
FEL ČVUT, Praha

Pavel Kordík
FIT ČVUT, Praha

Vojtěch Svátek
FIS VŠE v Prahe

Karol Matiaško
FRI ŽU, Žilina

Petr Šaloun
FEI VŠB-TU, Ostrava

Ján Paralič
FEI TU, Košice

Michal Valenta
ČVUT FIT, Praha

Organizing Committee

Chair

Ivan Srba, FIIT STU, Bratislava

Members

František Babič
FEI TU, Košice

Ondrej Kaššák
FIIT STU, Bratislava

Petr Šaloun
FEI VŠB-TU, Ostrava

Tibor Krajčovič
FIIT STU, Bratislava

Dušan Chlapek
FIS VŠE, Praha

Katarína Mršková
FIIT STU, Bratislava

Ľubica Jančaťová
FIIT STU, Bratislava

Jakub Ševcech
FIIT STU, Bratislava

Michal Kompan
FIIT STU, Bratislava

Marek Lóderer
FIIT STU, Bratislava

Obsah

| | |
|---|-----------|
| Pozvané prednášky | 1 |
| Vývoj databází a jeho reflexe v konferenciích DATASEM, DATAKON a Data a znalosti v letech 1981 - 2016 | |
| <i>Jaroslav Pokorný</i> | 3 |
| Skutočné potreby podnikov na zber a spracovanie externých dát – prípadové štúdie z praxe | |
| <i>Filip VÍTEK</i> | 15 |
| Vizualizácia informácií - aktuálne výzvy a trendy | |
| <i>Matej Novotný</i> | 27 |
| Graph Mining: Applications | |
| <i>Karel Vaculík</i> | 31 |
| Aktuálne dianie v oblasti otvorených údajov v SR (2016) | |
| <i>Peter Hanečák, Lubor Illek</i> | 35 |
| Aktuální dění v oblasti otevřených dat v ČR | |
| <i>Dušan Chlapek, Michal Kubáň</i> | 43 |
| Rozpoznání atributů vozidel pomocí metod strojového učení | |
| <i>Jan Sedláček</i> | 49 |
| Aplikácia prístupov hlbokého učenia na riešenie (ne)štandardných úloh strojového učenia | |
| <i>Michal Barla, Peter Lacko, Mária Šajgalik</i> | 53 |
| R vs. Python – which one fits you best? | |
| <i>Jakub Ševcech, Peter Laurinec, Ondrej Kaššák</i> | 59 |
| Analýza dát, dolovanie v dátach a strojové učenie | 63 |
| Course Similarity Analysis | |
| <i>Hana Bydžovská</i> | 65 |
| Detekcia nebezpečných aktivít v záznamoch udalostí mobilných zariadení | |
| <i>Štefan Dlugolinský, Giang Nguyen a Ladislav Hluchý</i> | 69 |
| Použitie transformačnej regresnej techniky pre dolovanie v údajoch | |
| <i>Peter Krammer, Ladislav Hluchý</i> | 75 |

| | |
|--|------------|
| Strojové učení pro analýzu rodinného podnikání <i>Juraj Michalik, Luboš Popelínský, Klára Antlová a Petra Rydvalová</i> | 81 |
| Anomaly detection for aircraft engine fault prediction <i>Tomáš Rudolecký</i> | 85 |
| Systém na podporu rozhodovania pomocou jednoduchého a efektívneho pochopenie medicínskych záznamov <i>Michal Vadovský, František Babič, Miroslava Muchová</i> | 89 |
| Analýza a spracovanie textu | 95 |
| Analýza článků z českých zpravodajských serverů <i>Markéta Filipová, Jaroslav Kuchař</i> | 97 |
| Interactive Evolution and Poem Models in Haiku Poetry Generation <i>Miroslava Hrešková, Kristína Machová</i> | 103 |
| Slovenský stemmer emocionálnych slov <i>Zuzana Nemčišinová, Martin Mikula, Kristína Machová</i> | 109 |
| Text Analyzing of Aviation Safety Reports <i>Lama Saeeda, Petr Křemen, Marek Štumper</i> | 115 |
| Hybridný prístup na klasifikáciu názorov <i>Katarína Simková, Martin Mikula, Kristína Machová</i> | 121 |
| Automatická anotácia a tvorba rečového korpusu prednášok TEDxSK a JumpSK <i>Ján Staš, Tomáš Kocúr, Peter Vízlay</i> | 127 |
| otazkovac: A Question Generator for Slovak Stories <i>Marek Šuppa, Marek Nagy</i> | 133 |
| Modelovanie používateľa, personalizovaný web, odporúčanie | 139 |
| Projekt HIBER: hlbšie poznávanie správania sa človeka v digitálnom priestore <i>Mária Bieliková, Pavol Návrat, Jakub Šimko, Jozef Tvarožek, Michal Barla, Róbert Móro, Eduard Kuric, Martin Labaj, Martin Konôpka</i> | 141 |
| Discovering User Preferences in Gamification for Libraries: A Methodological Approach <i>Andrea Hrečková</i> | 145 |
| Order Sensitive Measures of Preference Estimation Quality along Users <i>Michal Kopecky, Marta Vomlelova, Peter Vojtas</i> | 151 |

| | |
|---|------------|
| DevACTs: Zber a vyhodnocovanie aktivít, úloh a zdrojového kódu vývojárov <i>Karol Rástočný, Martin Konôpka, Mária Bieliková, Pavol Návrat</i> | 155 |
| Information behavior of researchers: contexts of digital scholarship <i>Jela Steinerová</i> | 159 |
| Prezentácia personalizovaných odporúčaní v prostredí webu <i>Martin Svrček, Michal Kompan</i> | 167 |
| Modelovanie informácií a znalostí, reprezentácia sémantiky | 171 |
| OpenPonk - platforma pro konceptuální modelování ve výuce, vědě a praxi <i>Jan Blizničenko, Peter Uhnák, Robert Pergl</i> | 173 |
| RDF Storage for Semantic Big Data Historian <i>Václav Jirkovský, Martin Possolt, Marek Obitko</i> | 179 |
| Modelování a transformace fiskálních datasetů technologiemi RDF v projektu OpenBudgets.eu <i>Jakub Klímek, Jindřich Mynarz, Vojtěch Svátek</i> | 183 |
| Grafová databáze jako úložiště metadat pro data lineage - zkušenosti a výzvy <i>Karel Quast, Michal Valenta</i> | 187 |
| Životní situace jako základní výchozí bod eGovernmentu <i>Václav Řepa</i> | 193 |
| Explorace společných charakteristik ontologií formální konceptuální analýzou <i>Ondřej Zamazal, Vojtěch Svátek</i> | 201 |
| Sociálny web a jeho aplikácie | 205 |
| Šírenie správ a vzťahy medzi užívateľmi v sociálnych sieťach <i>Lubomír Antoni, Stanislav Krajčí, Ondrej Krídlo</i> | 207 |
| Stance detection in online discussions <i>Peter Krejzl, Barbora Hourová, Josef Steinberger</i> | 211 |
| SoSIREČR – Sociální síť informatiků v regionech České republiky <i>Jaroslav Pokorný, Peter Vojtáš</i> | 215 |
| Twitter a #brexit sentiment <i>Petr Šaloun, Radka Cepláková</i> | 217 |

| | |
|--|------------|
| Aplikácie inteligentných znalostných technológií | 223 |
| Extrakcia štruktúrovaných objektov z webových portálov na pár klikov <i>Peter Gurský, Milan Vereščák</i> | 225 |
| Vizualizace dat s využitím frameworku AngularJS <i>Petr Kukrál, Martin Dostal, Dalibor Fiala</i> | 229 |
| EEG a rozpoznávanie výrazu tváre: Porovnanie prístupov na meranie emócií <i>Róbert Móro, Jakub Šimko, Peter Gašpar, Tomáš Matlovič, Jozef Tvarožek, Mária Bieliková</i> | 235 |
| Considerations about Data Processing, Machine Learning, HPC, Apache Spark and GPU <i>Giang Nguyen, Ján Astaloš, Ladislav Hluchý</i> | 241 |
| Podpora zdieľania znalostí vo vzdelávacích kurzoch prostredníctvom CQA systému Askalot <i>Ivan Srba, Mária Bieliková</i> | 249 |
| Návrh a implementácia aplikácie pre tvorbu prezenčných listín pomocou mobilného zariadenia s technológiou NFC <i>Zuzana Vantová, Vladimír Gašpar</i> | 253 |
| Využití cloudu pro dolování asocičních pravidel z velkých dat přes webové rozhraní <i>Václav Zeman, Stanislav Vojíř, Jaroslav Kuchař, Tomáš Kliegr</i> | 259 |
| Smerovanie dizertačných projektov | 265 |
| Web-Human Interaction Based on Ontology Query <i>Jana Ahmad, Petr Křemen</i> | 267 |
| miXGENE: an Effective Public Tool for Integrative Analysis of High-throughput Omics Data <i>Michael Anděl, Pavel Strnad, Jiří Kléma</i> | 271 |
| Využívanie hierarchie vetných vzorov pre automatizovanú tvorbu otázok <i>Miroslav Blšták, Viera Rozinajová</i> | 277 |
| Model používateľa pre jeho identifikáciu <i>Kamil Burda, Daniela Chudá</i> | 283 |
| Automation and visualization in ontological engineering leveraging on background models <i>Marek Dudáš, Vojtěch Svátek</i> | 287 |

| | |
|---|-----|
| Vplyv individuálnych vlastností používateľov na výsledky používateľských štúdií <i>Patrik Hlaváč, Mária Bielíková</i> | 291 |
| Metóda kombinácie predikčných modelov na spresnenie predikcie spotreby elektrickej energie <i>Marek Lóderer, Viera Rozinajová</i> | 295 |
| Analýza dát za účelom zlepšenia konkrétneho firemného procesu logistickej firmy <i>Miroslava Muchová, Ján Paralič</i> | 299 |
| Rozpoznanie podobnosti textov, programových kódov <i>Juraj Petrik, Daniela Chudá</i> | 305 |
| Automatické generovanie prediktorov a ich využitie pri dolovaní pravidiel <i>Michal Puheim, Kristína Machová, Ján Paralič</i> | 311 |
| Structural Features Extraction from Text using Applicative Supercombinator Form <i>Michal Sičák, Ján Kollár</i> | 317 |
| Hierarchické modelovanie témy nad prúdmi dát zo sociálnych sietí s využitím formálnej konceptovej analýzy <i>Miroslav Smatana, Peter Butka</i> | 321 |
| Cluster Based Symbolization <i>Milan Spišiak, Ján Kollár</i> | 325 |
| Predikcia Parkinsonovej choroby pomocou signálov reči použitím metód dolovania v dátach <i>Michal Vadovský, Ján Paralič</i> | 329 |

Pozvané prednášky

Vývoj databází a jeho reflexe v konferencích DATASEM, DATAKON a Data a znalosti v letech 1981 - 2016

Jaroslav Pokorný

MFF UK, Malostranské nám 25
Praha, Česká republika

`pokorny@ksi.mff.cuni.cz`

Abstrakt. Padesát let vývoje databází je již úctyhodné číslo. Se zhruba desetiletým zpožděním jsme ho zaznamenali i v Československu. Československá odborná komunita se zapojila rychle do této atraktivní problematiky. Historicky nejstarší odborná setkání pod názvem DATASEM (DATabázový SEMinář) započala již v r. 1981. Charakteristická pro tyto první konference byla velmi plodná symbióza odborníků teorie i praxe. Setkání se totiž hojně účastnili vedle akademiků i zástupci komerční sféry. Cílem článku je ukázat, jak se se světový vývoj databází odrážel a odráží v těchto odborných národních setkáních, tj. ve dvaceti letech semináře (později konference) DATASEM, pokračujícího dalších čtrnáct let jako DATAKON a konečně od r. 2015 jako konference Data a Znalosti.

Typ příspěvku: Zvaná přednáška

Klíčová slova: databáze, databázový systém, relační model dat, web, heterogenní datové zdroje, objektově-orientované databáze, objektově-relační databáze, ontologie, XML, NoSQL, NewSQL

1 Úvod

Postavení *databázových systémů* (DBS) v informatice se od počátku jejich vniku týkalo dvou základních problémů:

- jak efektivně ukládat data na vnějších pamětech,
- jak efektivně formulovat dotazy na takto uloženými daty.

Na ně se naroubovaly další problémy, jako je návrh databází, jejich začlenění do informačního systému (IS) organizace, integrace s daty na webu apod. Jak ukazuje historie, vývoj vedl vedle prací na vhodném software ruku v ruce s vývojem teorie databází, specializovaných časopisů, odborných konferencí a s výukou databází na většině škol zabývajících se informatikou.

Samostatné databázové konference započaly v Československu počátkem 80. let zřejmě seminářem DATASEM '81. Sborníky z tohoto semináře vydával dlouhá léta Dům techniky ČSVTS Praha. Jeden z prvních článků [11], kde by představen jazyk

Sequel (později SQL), však byl prezentován na dnes už kultovním semináři SOFSEM již v r. 1978.

Z dalších akcí věnujících se databázím stojí za zmínku konference Moderní databáze, která existovala od r. 1986 s několika pauzami až do r. 2012. Sloužil-li v prezentacích DATASEM a jeho další pokračování spíše akademické sféře navštěvované rovněž účastníky z komerční sféry, u Moderních databází to bylo obráceně. Hlavní přednášky zde byly od firem a společností prodávajících či vyvíjejících databázový software, přičemž přednášející z vysokých škol spíše ukazovali současné trendy a přehledy databázových technologií. Z dalších, vztažených konferencí můžeme jmenovat Objekty a Systémovou integraci, kde databáze vždy tvořily pouze část širší problematiky.

Zaměříme-li se na DATASEM, ten byl od 15. ročníku nazván konferencí. Od r. 2001 až do r. 2014 pokračoval po novém názvem DATAKON. Následující rok došlo k integraci konferencí DATAKON a Znalosti s novým názvem Data a znalosti. Připomeňme, že konference Znalosti existovala od r. 2001 a byla orientována zejména na reálně využitelné nástroje, datové zdroje a aplikace v oblasti znalostních technologií. Je však příznačné, že do r. 2001 se znalostní problematika spolu např. s expertními systémy objevovala i na pořadu DATAKONU (viz např. článek Báze znalostí a databáze z hlediska expertních systémů od P. Jirků v r. 1985 a další články od P. Bartoše a P. Hájka v r. 1986).

Připomeňme rovněž, že obě konference byly vždy organizovány ve spolupráci české a slovenské odborné komunity, v některých ročnících dokonce s mezinárodní účastí.

V r. 2005 bylo v [17] zdůrazněno, že DATAKON již není výlučně databázovou konferencí. To bylo přirozené. Stále více se smazávaly rozdíly mezi jednotlivými disciplínami. Dobrým příkladem je zpracování dat v prostředí webu, kde se potkávají databáze, logika, umělá inteligence, zpracování přirozeného jazyka a další obory. DATAKON se těmto trendům nevyhýbal. Podpora mezioborové komunikace je patrná i z posledního vývoje – vzniku konference Data a znalosti.

Cílem příspěvku je prezentovat populárně historii konferencí DATASEM a DATAKON, dále pak jejich poslední variantu Data a znalosti. Historický pohled byl v minulosti konference prezentován vícekrát – jednou v r. 2000 k příležitosti jejího 20. výročí [14], dále pak v r. 2005 [17], kdy jsme se dokonce pokoušeli objevit korelaci mezi obsahem těchto konferencí a databázovými trendy ve světě. Tento příspěvek nabízí po dalších 11 letech skromnější přehled těchto korelací bez hlubší analýzy příspěvků. V kap. 2 zmíníme něco z historie databází v Československu na pozadí s jejich vývojem ve světě zhruba do začátku 90. let. Kap. 3 vyzdvihuje dva základní směry v rozvoji databází 70. a 80. let – rozvoj SQL a vliv objektově-orientovaného programování. Kap. 4 popisuje stručně 90. léta z hlediska integrace objektů a tabulek v objektově-relačním modelu dat a integrace heterogenních dat. Zmíněn je i vztah databází a webu. Kap. 5 se věnuje přechodu do 3. tisíciletí s nástupem XML databází a postupnému vlivu fenoménu Big Data na databázovou technologii. V kap. 6 naznačíme některé směry vývoje databází a výhledy do budoucnosti konference Data a znalosti.

2 Historie databází v Československu

V [20] jsme uvedli výstižně formulované pravdy odborníka na SQL Joe Celko doplněním klasického citátu anglického básníka T.S. Eliota, který říká:

Kde je moudrost?
Ztracena ve znalostech.
Kde jsou znalosti?
Ztraceny v informacích.

A J. Celko pokračuje:

Kde jsou informace?
Ztraceny v datech.
Kde jsou data?
Ztracena v databázích.

Tyto citáty naznačují, že od dat ke znalostem, či dokonce k moudrosti je daleko. Databázové technologie se o to také ani nesnaží, tyto cíle jsou spíše vyhrazeny formalizaci znalostí, sémantickému popisu webu a rovněž umělé inteligenci.

Historie databází je dostatečně známá z řady databázových učebnic zejména těch zahraničních, jako jsou např. knihy [5], [21], [6], [22], či teoretičtější [8]. Ve střední a výhodní Evropě byl však jejich vývoj přeci jen specifitější. Bylo typické, že v Československu mělo využití DBS v praxi vždy trochu zpoždění. Stejně tomu bylo i na úrovni relevantních informací zvláště v akademické sféře, kde nedostatek odborné literatury a velmi omezená možnost návštěv zahraničních konferencí hrály svoji negativní roli. Nicméně první knihu o databázích od J. C. Date z r. 1976 jsme viděli poprvé v rusém vydání někdy v r. 1977. Překlad knihy Database systems od D. C. Tsichritzise a F. H. Lochovského napsané v r. 1977 však vyšel v Československu až v r. 1987 [24].

Vynecháme-li tzv. hromadné zpracování dat využívající přímo souborové techniky a indexování dat, má databázová technologie kořeny v *síťovém databázovém modelu*. V r. 1965 se formovala konference o jazycích datových systémů (Conference on Data Systems Languages - ve zkratce CODASYL). V rámci této konference byl vytvořen výbor známý jako Database Task Group (DBTG), který měl za úkol standardizačním postupem vytvořit koncepci *databázového systému* (DBS). Vznikaly *síťové systémy řízení bází dat* (SRBD) jako IDMS, u nás známý z éry sálových počítačů. Dokonce již od počátku 60. let byl pod vedením Ch. Bachmana vyvíjen SRBD IDS, který významně ovlivnil práci výboru DBTG. IDS je považován za první databázový software vůbec. V r. 1971 vydal výbor zprávu "The DBTG April 1971 Report", kde se objevily dnes dobře známé databázové pojmy jako *schéma databáze*, *jazyk pro definici schématu*, *subschéma* apod., jakož i celková architektura *síťového databázového systému*.

Téměř paralelně se vyvíjely hierarchické databáze využívající nikoliv speciální grafy typů záznamů ale pouze stromy (hierarchie). Na rozdíl od síťových nemají hierarchické databáze standard. Historie *hierarchického datového modelu* je nedílně spjata se SRBD IMS (Information Management System). Jak IDMS, tak IMS se používaly od konce 70. let i v Československu. Nahlédnutím do sborníku DATASEM '81 uvidíme, že se nereférovalo jen o IDMS, ale i na počítačích Siemens speciální variantě síťového modelu SESAM, či o domácím databázovém produktu s cobolskými datovými strukturami SOFIS (vyvinutém ve Výzkumném výpočtovém středisku v Bratislavě).

Je příznačné, že relační databáze se u nás objevují mnohem později, až v 80 letech. Od uvedení relačního modelu dat (RMD) E. F. Coddem [4] téměř 10 let trvalo, než se relační databázová technologie vyvinula natolik, aby byla z hlediska výkonu DBS v reálném prostředí srovnatelná s tehdejšími síťovými a hierarchickými protějšky. Připomeňme zde pionýrskou implementaci IBM ze 70. let, jako je System R (předchůdce dnešního SRBD DB2), nebo QBE z r. 1978. Třetí implementací byl INGRES z University of California. Z komerčních relačních produktů 80. let se mezi průkopnické řadí Oracle, Sybase, RDB (firmy DEC), Informix a Unify.

S rozvojem databázové technologie byly zákonitě vyvíjeny i přístupy k návrhu relační databáze. E. F. Codd totiž neřikal, jak navrhnout tabulky. Jeho cílem bylo, aby byly ve 3. normální formě, což nebylo pro návrháře analyzujícího danou aplikační doménu vždy jednoduché. Důležitým výsledkem byl vznik *E-R modelu*, který P. Chen publikoval v r. 1976 [3]. E-R model dával možnost konceptuálního modelování se systematickým přístupem k výslednému návrhu schémat relací. Přestože má E-R model mnoho odpůrců, je dnes tato koncepce ve svých četných variantách *de facto* standardem ve světě strukturovaných metodologií návrhu nejen databází, ale i obecnějších systémů. Kromě toho jsou na ní vybudovány i prostředky objektové.

Konceptuální modelování mělo hlubokou tradici i u nás, jak dokazují semináře DATAKON '81 a '82. Je zde prezentován *databázový model HIT* založený na jednoduché teorii typů a typovaném lambda kalkulu [25]. Funkcionální přístup HITu se v r. 1985 dokonce prosadil na významné databázové konferenci VLDB [26]. Kapitola o konceptuálním modelování [12] se dostala do překladu knihy [24]. Spolu s konceptuálním modelováním se rozvíjí i s funkční analýza, tj. SRBD se uvažuje v prostředí IS.

Rychleji se vyvíjela relační teorie, která dnes tvoří základ databázové teorie vůbec. Základní pojmy jako relační algebra a relační kalkul, normální formy, či teorie transakcí se v modifikované podobě dostávají i do dalších modelů. První detailnější domácí kniha o databázích vyšla v r. 1992 [13].

3 Relace a objekty

3.1 SQL

Tvůrce objektově-relační technologie ve firmě INFORMIX – M. Stonebraker kdysi prohlásil, že SQL je mezigalaktický dotazovací jazyk. O to více je to pravda dnes. Vše se přizpůsobuje SQL. Počátky jazyka SQL sahají do r. 1974, kdy se ještě nazývá Sequel a je zaměřen hlavně na svou dotazovací část. Jeho prototypová implementace byla součástí Systému R vyvíjeného v IBM v San Jose, kde byl zaměstnán i E. F. Codd.

Od prvního standardu zvaného SQL86 se na cestě vývoje SQL objevily milníky SQL89, SQL92, SQL:1999, SQL:2003, SQL:2006, SQL:2008 a SQL 2011. Všimněme si dvou věcí: mnohaleté vzdálenosti standardu z r. 1999 od standardu 1992. Je to dáno tím, že SQL založený na RMD do sebe absorboval objektové rozšíření. V letech 1999 – 2006 zase SQL vstřebával datový model XML. Zapomenout nesmíme rovněž na část standardu SQL/MM z r. 2003 obsahující rozšíření směrem k textům, prostorovým objektům, obrázkům a dolování dat.

Části standardu SQL jsou číslovány od 1 do 14. Za zmínku stojí, že části 5, 6, 8 neexistují, dočasně pozastaven je vývoj části 7 – SQL/Temporal (částečně implementován v ORACLE 11g, IBM DB2 pro operační systém z/OS, Teradata 13.10), zrušen byl vývoj části 12 – SQL/Replication. Zatím aktuální je standard SQL:2011, kde je třeba k dispozici příkaz pro „vypnutí“ integritních omezení. Obsahuje také podporu temporálních databází, která se ovšem liší od původního přístupu zrušené části 7.

3.2 Objektová orientace

Poněkud skromnější je historie *objektově-orientovaných* (OO) SŘBD koncipovaných v 2. polovině 80. let a identifikovaných Manifestem skupiny Altair v r. 1990. Byly inspirovány objektovým programováním a objektovými metodologiemi analýzy a návrhu. Nabízela se představa ukládat objekty do databáze a využít současně mnoha užitečných prvků OO technologie. Dalším důležitým důvodem pro použití OO byl fakt, že ne pro všechny aplikace byly relační SŘBD vhodné. Mezi reprezentativní příklady patří problémy modelování objektů v systémech pro návrh (např. CAD) či geografické IS. Živelnost vývoje OOSŘBD zastavil *de facto* standard ODMG-93 a jeho následné verze [2]. Byl přijat jak výrobci OOSŘBD, tak i tvůrci podpůrných nástrojů typu CASE pro návrh databází. Na DATAKONU o bylo o OOSŘBD detailněji referováno v r. 1992 (články J. Pokorného, A. Bicher a J. Valenty).

V současnosti existuje okolo 20 OOSŘBD¹. Přes počáteční optimismus se ukázalo, že počet nasazení OOSŘBD nerostl tak rychle, jak se předpokládalo. Také funkce a výkon těchto systémů jsou dosud na poměrně nízké úrovni. Řešení, které přijaly hlavně vůdčí relační databázové firmy v dalších letech, však tkví spíše v objektově-relačních SŘBD (ORSŘBD), které kombinují vlastnosti relačních SŘBD s přínosem OOSŘBD.

4 Databáze v 90. letech

90. léta se vyznačují snahami integrovat heterogenní data v podniku a rozšiřovat možnosti SŘBD o další datové typy. Jinými slovy řečeno, cílem bylo ukládat do databáze všechno, tj. možné i nemožné. Jedním ze směrů jak technicky vyřešit tyto problémy bylo rozšířit relační tabulky SQL o objekty. Výrobci SŘBD započali uvažovat netriviální rozsáhlé objekty typu text, audio, video atd. Pro tyto objekty bylo nutné vyvíjet nové dotazovací jazyky, které umožnily nejen nové typy dotazů (např. najdi k objektu v prostoru jeho nejbližšího souseda), ale i k přehodnocení dotazování jako takového. Vznikaly tzv. *univerzální severy s ad hoc* přidávanými novými datovými typy.

Integrace podnikových dat „ve velkém“ vedla k řadě architektur vycházejících z původních idejí distribuovaných databází řešených v 80. letech. Šlo vlastně o přístup zdola-nahoru k řešení distribuované databáze, založený na (ruční) integraci dílčích databázových schémat. Nemalé úsilí bylo věnováno řešení sémantických konfliktů mezi daty několika databází a transakcím nad více databázemi. Mimochodem, neúspěšnost těchto architektur v rámci IS podniku nakonec vedla k vývoji datových skladů (DW).

¹ https://en.wikipedia.org/wiki/Comparison_of_object_database_management_systems

Integrace dat se v DW provádí tak, že se potřebná data „vypumpují“ z operačních databází, vyčistí a uloží do databáze speciální.

4.1 Objekty, relace s objekty

Zlatým věkem OOSŘBD byla 90. léta (viz rovněž ročník *DATASEM* '93 a '94). Přestože existovaly názory, že OOSŘBD zcela vytlačí relační systémy, nestalo se tak. Naopak, relační systémy se přizpůsobily objektovým. Cílem bylo rozšířit relační datový model o objekty [23]. Vzniká *objektově-relační* (OR) databázový model reprezentovaný standardem SQL:1999.

OO a OR sice konvergují, ale spíše jen ve své části, která se týká dotazování. Pro OO i OR modelování je charakteristická především bohatost typů objektů, které jsou k dispozici, a rozšiřitelnost o další typy. Tradiční relační databáze umožňovaly modelovat takový svět jednoduše, ovšem za cenu mnohdy složitého a neefektivního přístupu k odpovídajícím datům.

ORSŘBD se pokoušejí překlenout mezeru mezi relační technologií a OOSŘBD. Přidávají možnosti ukládat objekty do relační databáze. Zapouzdřením metod a datových struktur může OR server vyvolat složité operace pro prohledávání a transformaci např. složitých multimediálních dat. Je tak vlastně řešen problém univerzálních serverů. Problémem ovšem vždy byla a je implementace takového přístupu. Nezapomeňme také, že relační funkčnost (dotazování, aktualizace apod.) je stále částí i ORSŘBD, tj. základními objekty jsou i nadále relace (tabulky).

ORSŘBD se zdály díky objektovému rozšíření SQL slibným článkem ve vývoji databázové technologie. Po více než 15 letech existence podobně jako dříve technologie OO však nedosáhly v aplikacích rozšíření srovnatelného s čistě relačními SŘBD.

4.2 Integrace heterogenních dat

Problém integrace heterogenních data se řeší v databázové historii neustále. 80. a 90. léta nabídla řadu technik, jak integrovat heterogenní data. Patří sem hlavně přístup přes globální schéma, federativní databáze a multidatabáze. Většina těchto systémů představovala statické řešení, které neobstojí v dynamickém prostředí, kdy jednotlivé databáze potřebné pro vyhodnocení nějakého požadavku nejsou ani dopředu známy.

Zřejmě nejméně statické řešení z těchto přístupů nabízela architektura federace. V prostředí internetu je však žádoucí integrovat i nestrukturovaná či semistrukturovaná data s dotazováním, které je založeno na volnějších principech, než např. pomocí SQL. Koncem 90. let se objevil nový datový model a jazyk XML pro popis semistrukturovaných dat. Jeho standard² se stal dalším milníkem na cestě databázovou historií.

S webovými službami se objevilo nepřesné vyhledávání, tak jak se používalo léta před tím v dokumentografických systémech (již v *DATASEM* '84). Uživatel chce vyhledat „podobné“ dokumenty, jako ten, který právě studuje, ale třeba také nějaké výrobce kočárků v jisté cenové kategorii, bez ohledu na to zdali budou všechny, tak jak

² <https://www.w3.org/TR/REC-xml/>

by mu je nabídnul relační systém založený na SQL, či dokonce seřazený podle nějakých uživatelských priorit.

Tab. 1 Kategorie příspěvků a jejich počty

| | 1981-2000 | 2001-2005 | 2006-2015 |
|---|-----------|-----------|-----------|
| Kategorie | # p | # p | # p |
| DB modely | 32 | 13 | 7 |
| Ontologie | NULL | NULL | 3 |
| NOSQL, Big Data | NULL | NULL | 7 |
| SŘBD cizí | 29 | 1 | 0 |
| SŘBD domácí | 17 | 0 | 0 |
| Distribované SŘBD | 22 | 0 | 4 |
| Teorie databází | 7 | 4 | 2 |
| Architektury DBS | 20 | 6 | 3 |
| Projektování IS | 70 | 12 | 14 |
| Dotazovací jazyky | 19 | 9 | 1 |
| Textové databáze, Zpracování textu na Webu | 20 | 5 | 14 |
| Sítě, Internet | 7 | NULL | NULL |
| Sítě | NULL | 3 | 0 |
| Web, XML, Open Data | NULL | 16 | 26 |
| Fyzické datové struktury, provoz DBS | 15 | 9 | 1 |
| Umělá inteligence | 34 | 9 | NULL |
| Dolování dat, Analytika | NULL | NULL | 13 |
| Aplikace | 15 | 12 | 16 |
| Přehledové příspěvky (Tutoriály) | 12 | 0 | 27 |
| Bezpečnost | 11 | 15 | 10 |
| Řízení IS/ITC | 17 | 3 | 10 |
| Ostatní | 20 | 6 | 24 |
| Celkem | 367 | 123 | 182 |

V tabulce 1 (části převzaty z [15], [14]) jsou počty příspěvků na konferencích DATASEM a DATAKON rozdělené podle témat, která byly jasně identifikovatelná a odrážela významné směry (samozřejmě ne všechny!) vývoje ve světě. Poslední sloupec již reprezentuje prvních 5 let 3. tisíciletí. NULL označuje, že kategorie není pro dané období definována.

5 Vstup do 3. tisíciletí

Přelom tisíciletí je i v databázové technologii ve znamení internetu a webu. Web poskytuje jednoduchý a univerzální standard pro výměnu informací. Po r. 2000 se intenzivně rozjely aktivity v technologii XML, zejména pak ve vývoji XML databází. Další

vývoj směřoval v souvislosti s rozvíjejícím se fenoménem Big Data směrem k tzv. *NoSQL databázím* (název byl pro tento software použit v r. 2009). Připomeňme, že pojem Big Data se známými charakteristikami se objevil v r. 2001 v [7]. Z pohledu webu se rozvíjel pojem sémantického webu jako databáze (viz např. článek Güttnera a Hrušky na DATAKON 2003) či články o RDF databázích z pozdějších let.

5.1 XML databáze

S příchodem XML vznikl nový databázový model, nové dotazovací či obecněji manipulační jazyky. Vznikají XML databáze.

Pro ukládání XML dat do databáze existují dvě základní architektury: databáze zpřístupňující XML data uložená např. v relačním SŘBD a nativní XML databáze. Nahlédneme-li do seznamu, který udržoval na svých webových stránkách R. Bourret [1] do r. 2010, zjistíme, že tehdy existovalo 24 komerčních produktů prvního druhu a 39 produktů druhého druhu.

V r. 2003 s objevila verze standardu SQL:2003 integrující datový model XML do relačního prostředí. V SQL:2006 dochází k úplné integraci XML do SQL včetně jazyka XQuery. Jazyk SQL rozšířený o XML se nazývá SQL/XML³. Tvoří část 14 standardu.

DATAKON reagoval na rozvoj XML databází článkem [16], který se stal jedním ze základních zdrojů pro českou knihu o XML technologiích [10] vydanou v r. 2008.

5.2 Směrem k velkým datům

Otočíme-li se směrem ke konkrétním problémům, které ovlivňují současné nové databázové technologie, existují dva zásadní - velikost databází a heterogenost databází.

V aplikacích se dostáváme do jednotek, jako jsou petabajt a exabajt. Reálný je i zetaabajt (10^{21}) v oblastech jako data z výzkumu Země či video-audio archivů.

Po mnoho let se ve vývoji IS spoléhalo na *vertikální škálování*, tj. investovalo se do nových a drahých velkých serverů. Bohužel, tento přístup použití architektury sdílení-ničeho vyžaduje vyšší úroveň dovedností a není v některých případech spolehlivý. Po přerozdělení dat za provozu může např. klesnout výkon systému. Rozdělování databáze mezi více (levných) strojů přidávaných dynamicky, tzv. *horizontální škálování*, může patrně zajistit škálovatelnost efektivněji a levněji. Než přizpůsobovat běžné SŘBD pro horizontální škálování, zdá se, že dnešní hojně citované NoSQL databáze navržené pro levný hardware a využívající rovněž architekturu sdílení-ničeho mohou být v některých případech řešením ještě lepším. Kromě cloud computingu se NoSQL databáze uplatňují v aplikacích Web. 2.0 a v sociálních sítích, kde horizontální škálování zahrnuje tisíce uzlů. Není náhoda, že nejvlivnější NoSQL databáze pocházejí z vývojových dílen firem Google a Amazon. DATAKON 2011 reagoval na tyto trendy v r. 2011 příspěvkem J. Pokorného [18] a v r. 2014 příspěvkem [19].

Významným ročníkem konference DATAKON byl DATAKON 2014 s tématy Big Data, Open Data, Linked Data. Zvané přednášky Big Data: jejich ukládání, zpracování a použití (J. Pokorný, MFF UK), Big Data zdaleka nejsou jen “velká” data (J. Slabý,

³ ISO/IEC 9075-14:2008: XML-Related Specifications (SQL/XML)

IBM) a Otevřená a propojitelná data (D. Chlapek, J. Kučera, FIS VŠE a M. Nečaský, MFF UK) rozvíjely detailně tato témata, samozřejmě ne pouze v databázovém kontextu. Z české odborné literatury pro oblast Big Data lze doporučit knihu [9].

Tab. 2 Nové databázové architektury v posledních 15 letech

| <i>Milník</i> | <i>Kategorie</i> | <i>Subkategorie</i> | <i>Reprezentanti</i> |
|---------------|------------------|-------------------------|---|
| 2009 | NoSQL | klíč-hodnota | Redis ⁴ |
| | | sloupcově-orientované | Cassandra ⁵ |
| | | dokumentově-orientované | MongoDB ⁶ |
| | | grafové databáze | Neo4j ⁷ |
| 2005 | BDMS | 1. Generace | Hadoop software stack |
| 2010 | | 2. Generace | Asterix software stack |
| 2011 | NewSQL | Obecné | NuoDB ⁸ , VoltDB ⁹ , Clustrix ¹⁰ |
| | | hybridy Google | Spanner ¹¹ |
| | | Hadoop-relační | Vertica ¹² , HadoopDB ¹³ |
| | | SQL-on-Hadoop | Hive ¹⁴ |
| | | NoSQL s ACID | FoundationDB, MarkLogic ¹⁵ , OrientDB ¹⁶ |

5.3 Nové databázové architektury

Samotné NoSQL databáze jsou sice vhodné pro určité aplikace využívající velká data, na druhé straně si však praxe postupně vyžádala složitější databázové architektury. Objevily se dokonce speciální SRBD nazývané v angličtině Big Data Management Systems (BDMS). Patří mezi ně zejména ASTERIX¹⁷ využívající speciální operace např. fuzzy spojení pro analytické účely. ASTERIX je součástí rozsáhlejšího softwarového zásobníku se vstupními body na různých úrovních pohledu na data, od těch nejvyšších (dotazovací jazyk AsterixQL), přes HiveQL, Piglet a další směrem k jobům v jazyku Pregel (slouží pro práci s grafy) a Hyracks na úrovni práce se soubory. Podobné snahy

⁴ <http://redis.io/>

⁵ <http://cassandra.apache.org/>

⁶ <https://www.mongodb.com/>

⁷ <https://neo4j.com/download/>

⁸ <http://www.nuodb.com/>

⁹ <https://voltdb.com/>

¹⁰ <http://www.clustrix.com/>

¹¹ <https://www.infoq.com/presentations/spanner-distributed-google>

¹² <http://www8.hp.com/us/en/software-solutions/advanced-sql-big-data-analytics/>

¹³ <http://db.cs.yale.edu/hadoopdb/hadoopdb.html>

¹⁴ <https://hive.apache.org/>

¹⁵ <http://www.marklogic.com/>

¹⁶ <http://orientdb.com/orientdb/>

¹⁷ <https://asterixdb.ics.uci.edu/>

jsou vidět i velkých databázových firem jako ORACLE. Např. Oracle Big Data Appliance kombinuje v SQL Hadoop a NoSQL v jeden dotaz SQL. V tabulce 2 jsou ukázány základní kategorie a subkategorie těchto nových databázových architektur.

Od r. 2011 se vyskytuje pojem *NewSQL* databáze. Jde o vysoce škálovatelní a elastické relační SŘBD, které

- jsou navrženy pro horizontální školování na strojích v režimu sdílení-ničeho,
- garantují ACID vlastnosti,
- aplikace na nich interagují s databází primárně přes SQL (včetně spojení),
- používají pro řízení souběžného zpracování protokol bez zamykání,
- poskytují vyšší výkon než tradiční relační.

Mezi *obecné* NewSQL patří ClustrixDB, NuoDB (vhodný pro cloudy) a VoltDB.

Zajímavá řešení architektur NewSQL jsou Spanner a F1 vyvinuté Googlem. Spanner používá hierarchie tabulek, které jsou semirelacemi, kde každý řádek má jméno (tj. vždy existuje primární klíč). F1 je SQL SŘBD vybudovaný nad Spanner.

Hadoop-relační hybridy zahrnují HadoopDB a Vertica. HadoopDB je paralelní databáze s Hadoop konektory transformující SQL dotazy do MapReduce jobů. Vertica je analytický SŘBD integrovaný s Hadoopem dvěma konektory umožňujícími vzájemný přenos dat mezi HDFS a systémem pomocí MapReduce. Do kategorie *SQL-on-Hadoop* patří např. Hive a jeho další varianty. Hive byl prvním SQL enginem na Hadoopu.

V podnikové sféře se objevují *NoSQL s ACID vlastnostmi*, někdy též nazývané *Enterprise NoSQL*. Tyto SŘBD zachovávají distribuovaný návrh, fault tolerance, jednoduché školování a jednoduchý, flexibilní databázový model. Co se týče transakčního zpracování, jde o CP (C – Consistency, P – Partition tolerance) systémy (tj. nezaručují obecně dostupnost) s globálními transakcemi. Patří sem např. FoundationDB, který je škálovatelným uložištěm typu klíč-hodnota, MarkLogic - dokumentově-orientovaná NoSQL databáze využívající pro ukládání dat formát JSON, HDFS, optimistické uzamykání. Distribuovaný SŘBD pro grafové databáze je OrientDB.

Tab. 3 Význačnost NoSQL ve světě databází z června 2016

| Pořadí | SŘBD | Databázový model | Skóre |
|--------|----------------------|-------------------------|---------|
| 1 | Oracle | relační | 1449.25 |
| 2 | MySQL | relační | 1370.13 |
| 3 | Microsoft SQL Server | relační | 1165.81 |
| 4 | MongoDB | dokumentově-orientovaný | 314.62 |
| 5 | PostgreSQL | relační | 306.60 |
| 6 | DB2 | relační | 188.57 |
| 7 | Cassandra | sloupcově-orientovaný | 131.12 |
| 8 | Microsoft Access | relační | 126.22 |
| 9 | SQLite | relační | 106.78 |
| 10 | Redis | klíč-hodnota | 104.49 |

V závěrečné tabulce 3 uvádíme část rozsáhlejší tabulky význačnosti databázových produktů z dobře udržovaného serveru DB-Engine¹⁸ (hodnotí 275 produktů). Vidíme, že NoSQL MongoDB, Cassandra a Redis se objevují v první desítce.

6 Závěr – aneb jak dál po r. 2015

V rozvoji databázových technologií se objevuje stále něco nového. Např. se hovoří o *Extreme Big Data* (EBD), tj. datech směřujících velikostí do YBajtů (10²⁴). Jak je ukládat a pracovat s nimi je jistě stále výzvou zejména na úrovni jejich distribuce a provozu v síti. Jiným problémem je, jak vybrat nějaký produkt či produkty do aplikační architektury s cílem integrace heterogenních (velkých) dat z různých zdrojů. Nabídka produktů s velmi odlišnými vlastnostmi je rozsáhlá a návrh a sestavení výsledné architektury vyžaduje velkou zkušenost a znalosti.

A co cloud computing? V souladu se současným vývojem ICT bychom mohli pokračovat ve stylu Eliota a Celka [20]:

Kde jsou databáze?

Ztraceny v cloudu.

Ano, databáze a ani architektura DBS nemusí být vidět. Pro uživatele to může být výhoda. Na druhé straně realizace efektivního cloudu opět je a stále bude výzvou. Obecně jde o Big Data, která jsou nejen velká, ale i heterogenní.

A naše konference? Rok 2015 znamenal v české a slovenské databázové komunitě posun v pojetí dvou konferencí do jedné nazvané Data a Znalosti. Integrace byla logickým vyústěním aktivit z r. 2013, kdy se poprvé konaly obě konference, i když zatím samostatně, na jednom místě. Na rozdíl od běžných odborných konferencí je její program založen pouze na zvaných přednáškách a posterech. V prvním ročníku konference bylo původně čistě databázové téma Big Data přirozeně doprovázeno tématy Big Analytics a pokročilá analytika. Ve zvaných přednáškách se uplatnila témata Řízení kvality dat s přihlédnutím k otevřeným a propojitelným datům (D. Chlapek, J. Kučera, FIS VŠE) a Vizualizace velkých dat (J. Géryk, L. Popelínský, FI MU).

Poděkování: Tato práce byla podpořena projektem P46 Univerzity Karlovy.

Literatura

1. Bourret, R.: XML Database Products. <http://www.rpbouret.com/xml/XMLDatabase-Prod.htm>, (2010).
2. Cattel, R.G.G., Barry, D.K. (Eds.): The Object Database Standard: ODMG 2.0. Morgan Kaufman Publishers, (1997).
3. Chen, P.: The Entity-Relationship Model – Toward a Unified View of Data. TODS, 1, 1, (1976).
4. Codd, E.F.: A Relational Model for Large Shared Data Banks. CACM, 13, 6, (1970).
5. Date, C. J.: An Introduction to Database Systems (8th Ed.), Pearson Education, (2003).
6. Elmasri, R., Navath, S.B.: Fundamentals of Database Systems (7th Ed.), Pearson, (2015).

¹⁸ <http://db-engines.com/en/ranking>

7. Laney, D.: 3D Data Management: Controlling Data Volume, Velocity and Variety. In: Application Delivery Strategies, Meta Group, (2001), 4 p.
8. Garcia-Molina, H., Ullman, J., Widom, J.: Database Systems - the Complete Book. Pearson Prentice Hall, (2009).
9. Holubová, I., Kosek, J., Minařík, K., Novák, D.: Big Data a NoSQL database. Grada 2015.
10. Mlýnková, I., Pokorný, J., Richta, K., Toman, K., Toman, V.: XML technologie. Principy a aplikace v praxi. Grada Publishing, a.s. Praha, (2008).
11. Pokorný, J., Benešovský, M., Krejčí, Fr.: Logika a dotazovací jazyky. Sbor. celost.sem. SOFSEM 78, Ždiar, (1978), pp. 201-235.
12. Pokorný, J.: Konceptuální modelování a jeho aplikace v D prostředí. Dodatek v [24], (1987).
13. Pokorný, J.: Databázové systémy a jejich použití v informačních systémech. ACADEMIA, (1992). 313 p.
14. Pokorný, J.: Odkud a kam kráčíte databáze (příspěvek k 20. výročí DATASEM). In: Proc. of 20th Annual Conf. DATASEM'2000, (J. Valenta ed.), Brno, (2000), pp. 85-104.
15. Pokorný, J.: Od Datasemu k Datakonu - aneb vývoj databází u nás. Soft. noviny, 4, (2001).
16. Pokorný, J.: XML databáze: současný stav a perspektivy. In: Proc. of the Annual Database Conf. DATAKON'2004, (K. Ježek ed.), Brno, MU Brno, (2004), pp. 161-181.
17. Pokorný, J.: DATAKON – aneb čtvrt století s databázemi. In: Proc. of the Annual Conf. DATASEM'2005, T. Hruška ed.), Brno, (2005), pp. 85-104.
18. Pokorný, J.: NoSQL databáze. In: Proc. of the Annual Database Conf. DATAKON'2011, (J. Vendulka, M. Rychlý eds.), Mikulov, October 15-18, VUT Brno, (2011), pp. 71-82.
19. Pokorný, J.: Big Data: jejich ukládání, zpracování a použití. In: Proc. of the Conf. DATAKON'2014, (J. Valenta ed.), Brno, (2014), pp. 85-104.
20. Pokorný, J., Valenta, M.: Databázové systémy. ČVUT v Praze, Česká technika - nakladatelství ČVUT, (2013), 274 p.
21. Ramakrishnan, R., Gehrke, J.: Database Management Systems, McGraw-Hill, (2002).
22. Silberschatz, A., Korth, H.F., Sudarshan, S.: Database System Concepts (6th Edition), McGraw-Hill, (2010).
23. Stonebraker, M., Brown, P.: Object-Relational DBMSs – Tracking The Next Great Wave. Morgan Kaufmann Publishers, (1999). (české vydání r. 2000 firma S&S).
24. Tsichritzis, D. C., Lochovsky, F. H.: Databázové systémy. Překlad: H. Tesařová, M. Benešovský, SNTL, (1987).
25. Zlatuška, J.: Databázový model HIT. In: Proc. DATASEM 81. (198), pp. 41-60.
26. Zlatuška, J.: Hit Data Model Data Bases from the Functional Point o For this first conference was characteristic f View. In: Proc. Of VLDB, (1985), pp. 470-477.

Annotation:

The database development and its reflection in conferences DATASEM, DATAKON and Data and Knowledge in years 1981 – 2016

Historically, the oldest professional meetings called DATASEM (DATAbase SEMinar) began in 1981. A very fruitful symbiosis of experts from theory and practice was characteristic for these first conferences. This paper aims to show how the world development of DB technology reflected and is reflected in these specialized national meetings, i.e. at twenty seminars (later conferences) DATASEM, continuing the next fourteen years as DATAKON and finally, since 2015, as the conference Data and Knowledge.

Skutočné potreby podnikov na zber a spracovanie externých dát - prípadové štúdie z praxe

Filip VÍTEK

Oddelenie CRM a BigData riešení
mediworx software solutions, a.s.
Einsteinova 19, 851 01 Bratislava, Slovenská republika

`filip.vitek@mediworx.sk`

Abstrakt. Využívanie externých dátových zdrojov, neraz Big Data charakteru, sa prehupla z roviny teoretických možností do prvých implementačných projektov aj v rámci stredoeurópskeho kontextu. V rámci príspevku autor sumarizuje ako správne odhaliť informačné potreby podnikov a ktoré úskalia pri ich zbere a spracovaní boli identifikované. Na sade konkrétnych príkladov zo SR a ČR ekonomického prostredia dokumentuje technologické a informačné požiadavky pre komerčné využitie tohto druhu služieb ako aj rôzne spôsoby monetizácie zbierania dát a ich použitia na biznis ciele. V závere príspevku autor naznačuje oblasti, v ktorých by výskumné a vzdelávacie inštitúcie mohli akcelerovať rozvoj tejto oblasti.

Typ príspevku: Pozvaná prednáška

Kľúčové slová: Big Data, komerčné využitie, informačné potreby, monetizácia údajov

1 Úvod

Technológie, umožňujúce pohodlne zbierať rozsiahle súbory dát z verejne dostupných serverov Internetu, sa stali dostupné pre rozsiahlu skupinu užívateľov do tej miery, že aj manažmenty komerčných spoločností a podnikov už nevnímajú *WebCrawling* ako sci-fi funkcionality alebo výskumné projekty čakajúce na zmysluplné speňaženie.

Iným dôležitým faktorom, ktorý akceleroval význam WebCrawlingu v komerčnom prostredí je skutočnosť, že pomocou výrazne spopularizovaných štatisticko-analytických softwarových nástrojov sa v období od roku 1995 tie najprogresívnejšie odvetvia (bankovníctvo, poisťovníctvo, telekomunikácie) dostali na horný limit informačného poznania zo svojich interných dát. Ak chcú podniky naďalej napredovať v informačnej výhode voči svojim konkurentom sú nútení obrátiť pozornosť k externým zdrojom dát.

Pre vzájomné pôsobenie viacerých (neskôr v texte objasnených) dôvodov, je priestor hromadného vyťažovania dát a Big Data¹ riešení atraktívnou nikou na pomedzí akademického výskumu a komerčnej sféry. Ako už to býva zvykom pre nástroje, ktoré sú „uprostred“, pre ich úspešné využitie je potrebné zohľadniť špecifiká oboch strán. Nasledujúce state príspevku pojednávajú o tom 1) ako správne rozpoznať príležitosti, kde komerčná sféra prejavuje dopyt po Big Data riešeníach, 2) aké nástroje sú najčastejšie pre realizáciu Big Data projektov realizované a 3) ako monetizovať/naceniť prínosy služby pre komerčný sektor. Prezentované princípy vznikli ako kultivácia projektových plánov jednotlivých prípadov využitia (use casov), ktorých realizáciu autor v rámci komerčného sektora v období rokov 2015-2016 aktívne manažoval alebo bol oboznámený s podstatou ich riešenia. Záverečné state príspevku autor adresuje vedeckým kruhom ako stimuláciu pre ďalšie rozvíjanie oblasti Big Data a dátových služieb ako aj dátových produktov.

2 Big Data nástroje - služby na pomedzí akademického výskumu a komerčnej sféry

Akademické aj politické špičky venujú téme intenzívnejšieho prepojenia vedecko-výskumnej práce v posledných rokoch veľmi intenzívnu pozornosť. Zo strany komerčného sektora sa uvedená diskusia odvíja najčastejšie smerom k potrebe zladenia akademického curricula s reálnymi pracovnými úlohami zamestnanca. Z pohľadu akademicko-kej obce sa najčastejšie pertraktuje možnosť pretaviť výskumné práce do komerčne využívaných projektov a tým zlepšiť financovanie vysokého školstva, zvýšiť prestíž výskumných teamov dosiahnutými praktickými aplikáciami ako aj motiváciu výskumných pracovníkov realizovať projekty, ktoré opustia laboratórne podmienky akadémie.

Miera prepojenia pretavenia výskumnej činnosti na komerčné využitie prirodzene kolíše od jedného vedného odboru k inému. Vo väčšine prípadov sa však realizuje buď:

- **PUSH princíp**, kde výskum ponúka pretlak nových prístupov a objavov (napr. aplikovaná matematika, materiálový výskum, ...). Častým sprievodným javom tohto prístupu je opatrnosť až skepsa zo strany komerčného sektora, keďže predstavované inovácie neraz požadujú radikálnu zmenu výrobných procesov alebo obsluhy koncového klienta. Tým pádom na pleciah akadémie zostáva marketing riešení.

ALEBO

- **PULL princíp**, kde naopak komerčný sektor aktívne vypisuje výzvy na vytvorenie produktových inovácií alebo nových metód. Spravidla však ide o oblasti, ktoré aj

¹ Aj keď tento pojem na seba v rôznych kruhoch preberá alternatívne podoby, pôvodná definícia tohto pojmu pochádza z publikácie META GROUP z roku 2001: <https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf> (viewed on 26.09.2016)

samotnou výskumnou obcou doposiaľ neboli plnohodnotne preskúmané. Kým prebehne prvotný výskum a experimentálne overenie hypotéz tak ubehnú neraz aj roky a komerčný sektor opäť zameriava svoju pozornosť na iné „horúce“ oblasti.

Z tohto pohľadu zaujímavým zistením je, že práve oblasť masových dátových zberov a ich analýzy (v tomto príspevku sumárne označovaná ako Big Data projekty) je unikátna vyváženosťou PULL a PUSH tendencií. S istou mierou odľahčenia možno konštatovať, že nielenže komerčný sektor ani vedecká obec sa nedokážu sami chopiť iniciatívy pre intenzívnejší rozvoj tejto témy (čo býva predmetom často aj iných vedných odborov), ale navyše v prípade vedných odborov podieľajúcich sa na Big Data projektoch tu nastala vzácna rovnováha medzi postavením oboch strán. Pre zdarné uchopenie tejto sľubnej príležitosti na spoluprácu preto našu pozornosť najprv upriamime na dôvody, ktoré predurčujú toto partnerstvo.

2.1 Dôvody partnerstva akadémie a komerčného sektora pri rozvoji Big Data

Napriek faktu, že koncept 3V, základný pilier Big Data rozvoja, bol kodifikovaný analytikmi META GROUP už v roku 2001, v stredoeurópskom priestore si téma Big Data aplikácii našla prvotné zhmotnenie v rozvojových projektoch komerčného sektoru až v období 2011-2015. Hlavným dôvodom pre inovačné spomalenie v predmetnej oblasti bola skutočnosť, že obchodné benefity plynúce z Big Data riešení sú najľahšie dosiahnuteľné u spoločností s veľkými portfóliami klientov (napr. Telekomunikácie, Bankovníctvo, Poistovníctvo, Energetika). V uvedených odvetviach však konkurenčné prostredie v stredoeurópskom regióne počas 2001 až 2009 sa ešte len rozvíjalo. Lokálni predstavitelia spomínaných odvetví teda neboli (na rozdiel od ich západných protažškov) nútení venovať intenzívnu pozornosť predmetnej téme a jej rozvoju.

Počas obdobia, keď zahraničné trhy začali experimentovať s Big Data prístupmi, ich lokálni predstavitelia stavili na *vyťaženie poznatkov interných dát*. Väčšina analytického úsilia bola venovaná do agregovania a dátových derivátov dát z vlastných procesov. Hoci tieto informačné

Okolnosti sa uviedli do pohybu po odznení celosvetovej finančnej krízy (2008-2010), keď lokálne spoločnosti boli konfrontované so skutočnosťou, že západní predstavitelia ich odvetví (s ktorými boli neraz i kapitálovo prepojení) už naskočili na vlnu Big Data riešení. Sľubné ekonomické výsledky prvotných Big Data projektov zlomili paradigmu, že Big Data je len teoreticky koncept bez ekonomického efektu². Ako sa ekonomika globalizuje, pre lokálnych veľkých predstaviteľov uvedených odvetví začalo narastať riziko, že online poskytované služby Big Data by *mohli dopomôcť k zintenzívneniu trhovej agresivity menších konkurentov* vedúcej k strate trhového podielu.

² V odborných kruhoch si Big Data dokonca vyslúžilo hanlivé prirovnanie k sexu medzi mladistvými: „Všetci o tom hovoria, aké je to skvelé, ale nik to v skutočnosti ešte nevyskúšal.“ Dan Ariely, In: Clau R., Big Data! Great! Now What? <http://www.slideshare.net/ricard-clau/big-data-great-now-what-symfonycon-2014>, 2014. (videné dňa 26.9.2016)

Ďalším podstatným faktorom pre vznik príležitosti pre Big Data partnerstvo je *technologická medzera*, ktorú so sebou implementácia Big Data projektov prináša. Pre rozmanitosť a rozsah dát nie je možné použiť tradičné SQL formáty ukladania dát, je potrebné siahnuť po moderných NoSQL³ formátoch úložísk a Open Source. Problémom na strane komerčného sektora však je, že dlhodobo sa spoliehal výlučne na štruktúrované úložiská a proprietárne software aplikácie na spracovanie a analýzu dát. Komerčná sféra teda dlhodobo nebudovala ľudské zdroje schopné pracovať s iným platformami. Naopak, akademické prostredie sa dlhodobo (aj z dôvodu licenčných nákladov) orientuje na Open Source software riešenia (R, Python, UNIX, MySQL, ...). Z uvedeného hľadiska teda možno uzavrieť, že akademické prostredie je (paradoxne) v oveľa priaznivejšie pripravené technologicky uchopiť Big Data projekty ako komerčná sféra.

Vyššie uvedená technologická medzera sa premieta aj do oblasti ľudských zdrojov. Keďže v stredoeurópskom priestore sa uvedenej téme venuje (aj vo vzdelávacom procese) pozornosť približne 5 rokov, vzdelávací systém ešte nestihol vyprodukovať dostatočný počet absolventov s patričnými zručnosťami. Komerčným subjektom, zažívajúcim *závažný nedostatok takto kvalifikovanej pracovnej sily*, nezostáva iné ako nakupovať drahé ľudské zdroje zo zahraničia alebo sa obracať na akadémiu a aktuálnych študentov vyšších ročníkov vysokých škôl.

Posledným, no nemenej dôležitým, dôvodom pre priaznivú klímu na spoluprácu akademického výskumu a komerčnej praxe v oblasti Big Data, je fakt, že väčšina Big Data projektov si vyžaduje *vývoj komplexných algoritmických konštrukcií* na vytváranie a následnú post-analýzu textov a iných neštruktúrovaných produktov. Vývoj týchto algoritmov si vyžaduje (neraz časovo náročne) systematické experimentovanie s rôznymi prístupmi. Akademické prostredie má možnosť vytážiť vysoký paralelizmus (ročníkové práce, viacčlenná výskumná skupina, etc.) pri testovaní jednotlivých prístupov, čo je značná výhoda oproti pomerne priamočiaremu procesu hľadania riešenia v komerčnom prostredí.

Súhra vyššie uvedených faktorov nahráva fakt, že akademické prostredie je v ČR/SR prostredí o mnoho viac pripravené uspieť vo vývoji Big Data riešení, ako komerčné prostredie, ktoré je však vystavenému náhlemu a intenzívnemu dopytu po Big Data riešení. Uvedené skutočnosti predurčujú oblasť Big Data (a s ním súvisiace vedné odbory) ako vhodný pre priestor pre partnerstvo vedeckého a komerčného sektora na vývoj riešení, kde sa rastúci dopyt zo strany komerčného sektora môže byť nasýtený relatívne vhodne pozicionovanými výskumnými kapacitami akadémie.

Detailnejšie skúmanie uvedených kľúčových faktorov atraktívnosti akademicko-komerčných partnerstiev v Big Data oblasti zakladá predpoklad, že v ich vplyve nastanú zásadné zmeny v horizonte 3-5 rokov. Hlavným aspektom v tom ohľade bude dostupnosť kvalifikovanej pracovnej sily pre Big Data záležitosti, ktorú vyriešia aj vlny absolventov prinášajúcich potrebné momentum na vytvorenie interných Big Data tímov v konkrétnych komerčných spoločnostiach. V horizonte 5 rokov následne možno

³ Definícia NoSQL prístupu a najpoužívanejších NoSQL databázových projektov pre Big Data projekty sumarizovaná na <http://nosql-database.org/> (videné dňa 26.9.2016)

predpokladať poklesu významu Big Data partnerstiev, ako túto úlohu preberú samostatné komerčné subjekty ako subdodávatelia Big Data služieb.

Identifikovaný potenciál na spoluprácu je možné pretaviť do skutočne úspešných projektov iba ak sa spoločné projektové úsilie bude pridržať dôležitých princípov pre komerčné využitie Big Data, ktorým sa detailne venuje najbližšia časť príspevku.

3 Reflektovanie potrieb komerčného v oblasti Big Data

3.1 Ako rozpoznať Big Data potreby komerčného sektora

Azda najdôležitejším aspektom úspešného Big Data partnerstva výskumných tímov a komerčného sektora je zacielenie na potreby komerčného sektora v tejto oblasti. Tak ako v iných prípadoch, komerčný sektor je ochotný priradiť hodnotu (a teda aj financovať) projekty, ktoré priamo vplyvajú na niektorú z *biznis priorít spoločnosti*. Vzhľadom na špecifickosť informačných prínosov Big Data projektov, ako najčastejšie sú požadované zo strany komerčného sektora nasledovné oblasti:

1. *Akvízia nových klientov*. Niektoré dátové stopy, ktoré nechávajú spotrebiteľia na webe pri používaní informačných portálov alebo sociálnych sietí zakladajú možnosť ich priameho oslovenia za účelom ponuky konkrétnej služby alebo tovaru. (napr. ak klient pridá do otvoreného, verejného profilu na sociálnej sieti informáciu, že hľadá radu pre výber destinácie pre dovolenku, môže byť hodnotným potenciálnym klientom pre cestovnú kanceláriu alebo portál rezervujúci lístky na prepravu). Pre Generovanie zoznamu perspektívnych klientov komerčný sektor označuje často aj ako Generovanie Leadov.
2. *Profilácia vlastných klientov*. Spoločnosti neraz disponujú len limitovanými informáciami o charaktere a životnom kontexte svojich klientov. Zozbieranie dodatočných verejných informácií o skupinách klientov (alebo priamo individuálnych klientoch) z diskusných fór, záujmových stránok alebo profilov sociálnych sietí tak môže byť cenným nástrojom pre zlepšenie cielenia marketingových aktivít klienta.
3. *Monitoring obchodného správania konkurencie*. Nemalá časť podnikania sa dnes už buď odohráva v online predajných kanáloch alebo aspoň v online prostredí prezentuje skladbu sortimentu, cenníky alebo iné cenné obchodné informácie. Systematickým zbieraním údajov o konkurentoch prináša danému komerčnému subjektu informačnú výhodu, ktorú vie pretaviť do dodatočných tržieb alebo úspory nákladov.
4. *Profilácia externých obchodných partnerov*. Tak ako možno vo verejných častiach webu zbierať informácie o koncových zákazníkoch podniku, je možné profilovať externými údajmi o dodávateľoch alebo iných dôležitých obchodných partnerov. Zdrojom dát pre tento druh Big Data projektov zväčša bývajú verejné registre, odvetvové stránky alebo webové sídla samotných profilovaných subjektov.
5. *Prevencia odchodu klientov*. Tak ako je pre podnik dôležité získavať nových klientov, je biznisovo hodnotné udržať si už získaných klientov. Než sa klient rozhodne zmeniť dodávateľa svojich služieb alebo tovaru, zvykne najprv preskúmať ponuky konkurenčných spoločností, prípadne využije služby cenového porovnávača. Preto

detegovanie interakcie klienta s konkurenciou môže byť pre komerčný subjekt hodnotným impulzom pre komunikáciu o vhodnosti aktuálnej formy produktov a služieb voči danému klientovi. Vhodnými dátovými zdrojmi pre tento druh projektov bývajú sociálne profily alebo web stránky konkurencie alebo dáta z cenových porovnávacích služieb.

Výber konkrétneho prínosu je zväčša determinovaný aj dostupnosťou dátových zdrojov pre niektoré z uvedených biznis motivátorov. Spoločnosti často preferujú promptne dostupné neúplné dáta s úmyslom získať informačnú výhodu a až následne doladiť precíznosť dátových odporúčaní. Pred spustením konkrétneho partnerstva sa odporúča preskúmať dostupnosť (a dátovú kvalitu) pre každý z biznis prínosov, keďže niektoré dátové zdroje môžu poskytovať vstupy aj pre niekoľko cieľov súčasne.

3.2 Najčastejšie technológie a nástroje pre riešenie komerčných projektov

Každý seba lepší projektový plán je náčrtom pokiaľ nie sú dostupné potrebné nástroje na realizáciu daných funkcionalít. Za pôsobenie autora v teame špecializujúcom sa na Big Data projekty možno skonštatovať, že najčastejšími požadovanými nástrojmi komerčných Big Data projektov sú nasledovné:

1. *Web crawling, masový zber externých dát.* Absolútna väčšina projektov ako jeden z prvých krokov vyžadovala vybudovanie automatizovaného robota na zber dát z definovaných zdrojov. Z technologické hľadiska zväčša išlo mini aplikácie v Python, Java Script alebo inom na web orientovanom algoritmickom jazyku. Pre účely využitia paralelizmu bude nutné pre výskumné teamy zvládať virtualizáciu serverov ako aj orchestráciu väčšieho počtu serverov pomocou nástrojov ako Chronos.
2. *No SQL úložisko zozbieraných dát.* Pri zbieraní externých dát má väčšina výstupov povahu textu, obrázkov, audio alebo video stopy. Pre efektívne uchovávanie týchto dátových formátov nepostačujú štruktúrované (SQL) databázové úložiska a je potrebné poskytnúť NoSQL úložiska. Z technologického hľadiska ide primárne o Open source produkty a riešenia (pre elimináciu licenčných nákladov) ako Cassandra, Mongo DB alebo HDFS.
3. *Text mining algoritmy.* Keďže najčastejším výstupom web crawlingu sú textové polia, azda najdôležitejším analytickom komponentom sú nástroje na dekompozíciu textu, jeho tagovania a následne párovanie a vzájomná súvislosť textových reťazcov. Hoci základné balíky textových analýz (stemming, TF-IDF, Ngrams, ...) poskytnú vhodný začiatok pre väčšinu úloh, ale neskôr bude potrebné programovať aj individuálne rutiny pre optimalizáciu (hlavne) párovacích algoritmov.
4. *Sémantická analýza.* Popri samotnej funkčnej analýze textových reťazcov sa často objavuje požiadavka na sémantickú/emotívnu analýzu dát, prisudzujúcu jednotlivým častiam textu odtienok nálady alebo postoja klienta.
5. *Machine learning algoritmy.* Popri analýze textu je druhou najčastejšou úlohou v Big Data projektoch Asociačná analýza (rule generation). Pre štruktúrované dáta sú často požadovanými komponentami autonómne, machine learning algoritmy pre klasifikáciu objektov alebo pre výpočty pravdepodobnosti (rozhodovacie stromy,

neurónové siete). Druhá skupina zmieňovaných algoritmov si však vyžaduje aj detailnú znalosť biznis prostredia a preto zriedkavejšie býva ponúkaná do externých partnerstiev na Big Data projekty.

6. *Špecifické non-text analýzy.* Sofistikovanejšie Big Data projekty zväčša prekonajú prizmu textových analýz a budú požadovať komplexnejšie algoritmy ako detegovanie patternov v obrázkoch, ich vzájomné stotožňovanie, prípadne dekompozícia audio alebo video stôp, prípadne párovanie časovo určených dát s video stopami.

Hoci akademické výskumné teamy prišli nepochybne do kontaktu aj so sofistikovanejšími formami analytických rutín a v mnohých z progresívnych vetiev prebieha intenzívny primárny výskum, pre podporu Big Data partnerstiev by sa výskumné teamy mali zamerať na zvládnutie a kapacitné vystuženie vyššie uvedených 6 oblastí Big Data.

3.3 Spôsoby monetizácie dátových analýz v Big Data prostredí

Popri technologických zmenách a špecifických oblastiach komerčného prínosu Big Data projektov si projekty v tejto oblasti vyžadujú aj implementáciu nových spôsobov monetizácie (speňaženia) výstupov analytického prostredia. Nasledujúca stať popisuje základné princípy **4 najčastejších spôsobov** monetizácie Big Data výstupov. Voľba vhodného monetizačného modelu je dôležitým predpokladom

(Jednorazová) Výskumná úloha. Pokiaľ predmetom analýzy je konkrétny časový rez dát alebo sú hromadne spracovávané historické, prípadne v čase stabilné dátové vstupy, je vhodné monetizovať výsledky formou výskumnej úlohy. Pre väčšinu spolupracujúcich komerčných subjektov sa tak akceptovateľná cena projektu pre výskumnú úlohu odvíja od alternatívnych interných nákladov nutných pre realizáciu danej úlohy. Pre kalkuláciu na strane akadémie je možné použiť hodinové (resp. manday) sadzby bežné v IT sektore podporujúcim daného koncového užívateľa dátových výstupov.

Opakujúca sa dátová služba. Ak je cieľom projektu poskytovať mapovanie dynamických dát, ich zmeny v čase alebo identifikovanie zmien v stave objektov alebo súvisiacich textov, nie je vhodné monetizovať výstupy pomocou jednorazových výskumných úloh, ale postaviť projekt ako dodávku kontinuálnej služby (zberu a analýzy dát). Nastavenie ceny dátovej služby je však potrebné už relativizovať k biznis prínosom, ktoré má služba v čase prinášať (napr. objem dodatočných tržieb alebo uspokojených nákladov). Potrebné je zohľadniť aj *ziskovú maržu daného odvetvia*⁴. Služba je zvyčajne dodávaná vo forme mesačných, kvartálnych alebo ročných poplatkov za používanie služby. Pre zodpovedné nacenenie je potrebné zohľadniť fakt, že kontinuálna povaha služby už si bude vyžadovať servisovanie (monitorovanie dostupnosti služby pre koncového klienta a korekcie algoritmov v čase).

Dátový produkt. Pokiaľ dáta prechádzajú aj vyššími úrovňami spracovania (napríklad analýza štatistického rozdelenia zbieraných veličín), je možné dodať výstupy analýzy aj ako dátový produkt. Na rozdiel od dátovej služby, v prípade dátového produktu

⁴ Údaje zverejňované štatistickým úradom alebo na portáli FinStat, vid': <https://www.fin-stat.sk/analyzy/financne-ukazovatele-slovenskych-firiem> (videné dňa 26.9.2016)

výskumný team neposkytuje komerčnému subjektu prístup k dátam, ale predáva ucelenú dátovú sadu („dátovú kocku“), zahrňujúcu všetky možné hodnoty (napr. hodnoty všetkých krvných tlakov v populácii). Dátový produkt je vhodnou formou monetizácie, ak sa predpokladá, že koncový užívateľ tohto dátového produktu potrebuje integrovať dátovú kocku do svojich systémov alebo ju bude ďalej v inej forme predávať svojim klientom. Pre necenenie dátového produktu sa zvyčajne používa ekvivalent niekoľko ročného používania dátovej služby, ktorá by poskytla rovnaké údaje. Ako vhodné obdobie pre výpočet je zvoliť obdobie, po ktoré sa predpokladá akceptovateľná miera aktuálnosti daných napočítaných dát.

Licencia na algoritmus. Na rozdiel od predaja samotných dátových výstupov, môže sa výskumný team rozhodnúť monetizovať práva na využívanie Big Data algoritmu. Tento - v našom regióne zatiaľ menej používaný spôsob monetizácie – môže byť zdrojom zaujímavých finančných prostriedkov pre výskumný team aj publicity v expertnej komunite. Je však dôležité poznamenať, že na to, aby mohol byť algoritmus predávaný ako nehmotný majetok (duševné vlastníctvo) musí byť chránený patentom alebo licenčnými zmluvami, ktoré výrazne navyšujú náklady na spustenie spolupráce. Zároveň tento spôsob monetizácie je možné použiť iba voči partnerom, ktorí sú schopní vo vlastnej infraštruktúre spúšťať a servisovať licencovaný algoritmus, ako aj odpisovať náklady ako dlhodobý nehmotný majetok (spravidla do 5 rokov odpisov). Uvedené špecifiká výrazne limitujú aj zoznam potenciálnych partnerov pre tento druh monetizácie.

Vo všeobecnosti možno povedať, že väčšina Big Data partnerstiev sa opiera o monetizáciu buď v podobe *Dátovej služby*, prípadne vo forme *Dátového produktu*. Monetizácia formou *Výskumnej úlohy* je síce univerzálnou formou, použiteľnou takmer pre ľubovoľný Big Data projekt, je však najmenej výhodnou formou monetizácie pre realizátora zberu a analýzy dát. Preto by nemala predstavovať prvú voľbu pre partnerstvá.

3.4 Oslovenie vhodného komerčného partnera pre Big Data projekt

Po tom, čo sme sa oboznámili s biznis prínosmi a najčastejšie požadovanými komponentmi, je dôležité ešte rozvážiť akým spôsobom sa vedecké teamy pokúsia o nadviazanie partnerstiev s potenciálnymi odberateľmi spomedzi radov komerčného sektora. V priestore ČR a SR Big Data oblasti možno odporučiť pre hľadanie nasledujúce dva prístupy:

- **Projekt za pomoci IT dodávateľa.** Pre väčšinu projektov je potrebné získané výstupy Big Data funkcionality ešte vsadiť do prezentačnej vrstvy (napríklad BI portál) alebo zintegrovat' s niektorou s existujúcou aplikáciou koncového zákazníka. Výskumný team zriedka disponuje zdrojmi, ktoré by systematicky mohli pracovať aj v priestoroch koncového zákazníka, preto (okrem prípadov uvedených v nižšom bode) je zrejme najvhodnejšou formou spolupráce realizovať Big Data projekt ako subdodávateľ niektorého z IT dodávateľov koncového zákazníka. Výhodou tohto prístupu je ak fakt, že IT dodávatelia majú zväčša rozvinuté nástroje marketingu IT služieb aj nad rámec pilotných koncových zákazníkov.

- **Priamy projekt s koncovým odberateľom.** Pri naplnení určitých podmienok je však napriek tomu možné pre výskumný team pracovať priamo s koncovým odberateľom. Vhodnými subjektmi na priamu spoluprácu sú spoločnosti, ktorých podstatu podnikania sa odvíja od IT produktov/služieb (prevádzkovatelia portálov, aplikácií alebo IT dodávateľia), prípadne významná časť ich podnikania je v online priestore (elektronické obchody, agregátory a rezervačné portály, ...). u týchto subjektov možno predpokladať, že ich miera porozumenia IT trendom bude dostatočná, aby vedeli byť partnerom pre expertnú špecifikáciu nástrojov a cieľov Big Dáta projektov. V ostatných prípadoch je vhodné použiť IT firmu ako prostredníka.

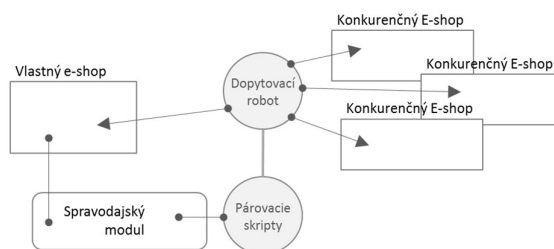
Z hľadiska všeobecných charakteristík vhodných subjektov pre Big Data projekty je možné špecifikovať nasledovné faktory zvyšujúce pravdepodobnosť záujmu o Big Data služby:

1. Veľký počet koncových klientov, dodávateľov alebo partnerov
2. Silné konkurenčné prostredie v odvetví
3. Multi-channel obsluha koncového klienta alebo predaj produktov
4. Vyššia frekvencia transakcií (nákupov) koncových klientov

4 Reálne Big Data projekty realizované v ČR a SR priestore

Dokumentované princípy a odporúčania pre sféru výskumno-komerčného partnerstva v oblasti Big Data, ktoré sú predmetom ostatných kapitol tohto príspevku, vnikli postupne ako destilácia opakujúcich sa princípov a predpokladov reálne implementovaných projektov v komerčnej sfére. Hoci pre nastavenie spolupráce s komerčnou sférou sú dôležité primárne uvedené princípy, pre lepšie pochopenie postojov komerčnej praxe je hodnotné vidieť uvedené pravidlá zasadené priamo v konkrétnych príkladoch. Detailný popis jednotlivých projektov je obsahovo nad rámec tohto príspevku. Autor sa preto spoľahol na kondenzovaný popis základných cieľov a výskumných metód p. Na záver každého z projektov je uvedená tá podmnožina identifikovaných princípov spolupráce vedeckých pracovníkov, ktorá bola obzvlášť dôležitá v danom projekte.

4.1 Prípady štúdiá: „Cenový monitoring konkurencie“



Obr. 1. Schéma dopytovania a analyzovania

Odvetvie: Online predaj kozmetiky a výživových doplnkov

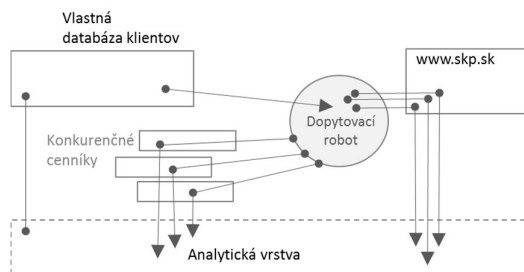
Ciele projektu: Na dennej báze monitorovať a reportovať správanie konkurenčných portálov predávajúci obdobný sortiment. Identifikovať prekryv produktovej ponuky s domovským e-shopom. Hlásiť akékoľvek zmeny v cenovej politike niektorého z portálov ako aj zalistovanie nových produktov konkurenciou.

Použité techniky: Automatický web-crawling, Text mining, Testy na zhody textových reťazcov, Analýzy časových radov

Najobťažnejšie elementy: Párovanie textových reťazcov názvov produktov (väčšina portálov nepoužívala žiadne unikátne ID produktov) a presné párovanie produktov bolo pomerne komplikované, pretože tie isté produkty sú vyrábané vo veľkom množstve podobných mutácií. Navyše z toho istého produktu sú vyrábané rovnaké balenia s rozličnou účinnou látkou alebo rozličným objemom balenia. Štandardné textové párovacie algoritmy vykazujú vysokú mieru false positive párovaní.

Princípy: *Hodnota pre klienta* = monitoring konkurencie, cenové spravodajstvo. *Forma monetizácie* = kontinuálna dátová služba, pravidelný (mesačný) poplatok za používanie služby. *Dodatočné komplikácie* = Paralelizmus pre maskovanie frekvencie dopytovania sa na konkurenčných portáloch

4.2 Prípadová štúdia: „Monitorovanie prepoistenia klientov“



Obr. 2. Schéma dopytovania pre monitorovanie prepoistenia klientov

Odvetvie: Komerčné poistenie majetku – poistenie vozidiel

Ciele projektu: Majitelia motorových vozidiel majú raz ročne (na výročný deň zmluvy) možnosť vypovedať zmluvu o Povinnom zmluvnom poistení (PZP) svojho vozidla. Keďže poistiť si svoje motorové vozidlo je podľa zákonnú povinné, každý klient, ktorý vypovie zmluvu sa musí pod hrozbou pokuty od úradov poistiť bezodkladne poistiť u inej poisťovne. Preto každý stratený PZP klient je automaticky klientom inej poisťovne. Na portáli www.skp.sk je dostupná služba, v ktorej si ako účastník dopravnej nehody môžete overiť, v ktorej poisťovni je poistené auto, s ktorým ste mali nehodu. (podľa ŠPZ daného auta).

Systematickým dopytovaním na ŠPZ odídených klientov je možné zistiť, ktorá poisťovňa ich odlákala. Systematickým dopytovaním na celý trh je možné dosiahnuť trhové spravodajstvo o tom ako sa zmenilo portfólio konkurentov alebo z akej konkurenčnej

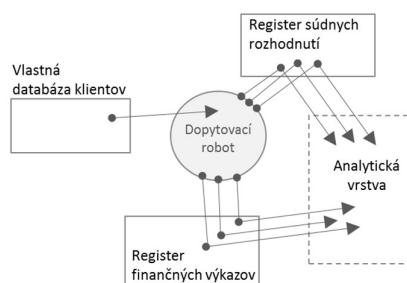
poisťovne sa podarilo nám odlákať klientov. Porovnaním cenových hladín a pohybov klientov je možné testovať cenovú elasticitu klientov pri zmene poistenia.

Použité techniky: Automatický web-crawling, Text mining, Algoritmus na efektívne dopytovanie celého trhu vozidiel.

Najobťažnejšie elementy: Text mining detailných popisov vozidiel a ich párovanie na cenníky konkurencie. Extrakcia cenníkov konkurenčných cenníkov poisťovní (vysoké množstvo kombinácií taríf).

Princípy: *Hodnota pre klienta* = monitoring konkurencie.
Forma monetizácie = analytická výskumná úloha

4.3 Prípadová štúdia: „Finančné spravodajstvo o platiteľoch poistenia“



Obr. 3. Schéma dopytovania pre Finančné spravodajstvo o platiteľoch poistenia

Odvetvie: Poistenie osôb hradené pre zamestnancov zo strany zamestnávateľa

Ciele projektu: Pri poistení, ktorého platiteľom poistného sú právnické osoby, je kľúčové monitorovať finančné zdravie platiteľov poistného. Poistné je totiž často vnímané ako jedna z prvých nákladových položiek, ktorej sa spoločnosti vo finančnej núdzi rozhodnú vzdať. Cieľom projektu bolo identifikovať právnické osoby, ktoré sa pravdepodobne dostanú do finančných problémov. Ako zdrojové dáta boli použité verejné registre súdnych rozhodnutí a registre finančných výkazov

Použité techniky: Automatický web-crawling, Text mining, Sémantická analýza

Najobťažnejšie elementy: Extrakcia kontextu súdnych rozhodnutí pre pochopenie kontextu a závažnosti rozhodnutia pre dopad na financie daného subjektu.

Princípy: *Hodnota pre klienta* = profilácia vlastných klientov. Forma monetizácie = kontinuálna dátová služba, pravidelný (ročný) poplatok za používanie služby

5 Záver

Viaceré trhové faktory spôsobili, že oblasť vedných odborov dotýkajúcich sa Big Data oblasti je priaznivo pozicionovaná na zakladanie partnerstiev s komerčným sektorom na vývoj Big Data projektov. Pre skutočné vytlačenie uvedeného potenciálu je potrebné

dôsledne poznať hlavné prúdy dopytu komerčnej sféry po predmetných službách ako aj sadu nástrojov, ktoré umožnia zdarnú realizáciu vytýčených Big Data projektov. Príspevok okrem pomenovania vyššie uvedených princípov projektov dopĺňa aj vymedzenie najčastejších modelov monetizácie Big Data výstupov, ako aj odporúčanie pre hľadanie vhodných partnerov z komerčného sektora. Všetky menované prvky sú následne ilustrované v sérii prípadových štúdií realizovaných v ČR a SR podnikateľskom prostredí.

Pri aplikovaní predstavených princípov, výskumné pracoviská majú unikátnu šancu pozdvihnúť mieru integrácie akademickej výskumnej činnosti s reálnou biznis praxou, ktorá je predmetom vízií ako akademickej tak i politickej reprezentácie. Kľúčové faktory zvyšujúce príťažlivosť spolupráce s komerčným v Big Data oblasti majú predpoklad zotrvať v platnosti v horizonte 3-5 rokov, po uplynutí ktorých je pravdepodobné predpokladať masové rozšírenie Big Data služieb príslušnými odvetvami samotného biznis prostredia.

Literatúra

1. Tsipstsis, K., Chorianopoulos, A.: Data Mining Techniques in CRM. John Wiley & Sons Ltd., Chichester UK, (2009).
2. Russel, M.A.: Mining the Social Web. O'Reilly Media, Inc., Sebastopol California, (2011).
3. Chapman, C., McDonnell Feit, E.: R for Marketing Research and Analytics. Springer International Publishing, Switzerland, (2015).
4. Grover, M., Malaska, T., Seidman, J., Shapira, G.: Hadoop Application Architectures: Designing Real-World Big Data Applications. O'Reilly Media, Inc., Sebastopol California, (2015).
5. Bahga, A., Madiseti, V.: Big Data Science & Analytics, A Hands-On Approach. Published by authors, (2016)
6. Liu B.: Web Data Mining. Springer-Verlag, Berlin, (2011).

Annotation:

Real needs of business entities in mass collecting and processing of external data – use cases of real CEE projects

Topic of external data, often Big Data like, usage has matured also in CEE region from theoretical concepts to first successful implementation projects. Author has taken part in several projects aimed at mass crawling of the external data for business sector. Following article summarizes most common informational needs of CEE businesses and principles necessary to withhold in forging successful Big Data partnerships between academic research bodies and business entities.

Article depicts technological and data requirements through set of real projects realized in CEE region within 2014-2016 period. Separate attention is dedicated to competition monitoring and collaborative filtering as the most common use cases of Big Data commercial projects recently. In the final part author indicates how to monetize Big Data services and products and indicates areas where universities can accelerate Big Data implementation wave.

Vizualizácia informácií - aktuálne výzvy a trendy

Matej Novotný

VIS GRAVIS, s.r.o.

novotny@visgravis.sk

Abstrakt. Interaktívna vizualizácia dát je dôležitou súčasťou analytických procesov, kedy úspešne prepája výpočtový výkon strojov s inteligenciou a skúsenosťami ľudí. Radikálne zmeny (Big Data, údaje v reálnom čase, bezpečnostná politika, sociálne siete atď.) však predstavujú nové výzvy aj pre overené postupy a musíme na ne vedieť reagovať. Sme preto svedkami evolučných aj revolučných zmien v oblasti vizualizácie. Príspevok predstaví tieto súčasné trendy, pomenuje nové výzvy v oblasti interaktívnej vizuálnej analýzy dát a načrtne atraktívne smery pre budúci výskum.

Typ príspevku: Pozvaná prednáška

Kľúčové slová: vizualizácia dát, interakcia, heterogénne dáta

1 Informácie v protismere

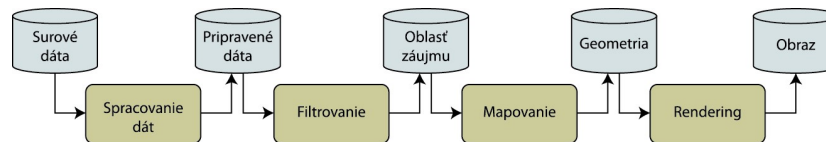
Sme svedkami radikálnych zmien vo svete informácií. Informačná superdiaľnica, ktorá sa objavila na prelome tisícročí s nástupom internetových technológií, bola v začiatkoch – použijúc dopravnú analógiu – iba jednosmerkou. Distribuovala dáta uložené v digitálnych obdobiach analógových knižníc k čitateľom. Bola to síce revolúcia v prístupe k informáciám, porovnateľná s vynálezom kníhtlače, ale skutočná zmena paradigmy prišla až v súčasnej dobe. Otvorením opačného smeru na informačnej superdiaľnici.

Big Data, sociálne siete, dátová žurnalistika, GPS, Internet vecí... to je len niekoľko pojmov, ktoré zmenili to, ako pristupujeme k dátam, ako ich spracúvame, analyzujeme a akým spôsobom z nich získavame poznatky o svete.

Dáta, s ktorými dnes pracujeme, rastú v objeme vďaka technológiám pre ich zber a uchovávanie. Integráciou dát z rôznych zdrojov vzniká heterogénne prostredie vyznačujúce sa veľkým rozsahom dát, ich veľkou variabilitou a narastajúcou dôležitosťou metadát. Zároveň vyvstáva otázka spoľahlivosti dát, ktoré vznikajú mimo kontrolovaného laboratórneho prostredia, v divočine digitálneho sveta.

2 Horúca dátová pôda

Mohlo by sa zdať, že pre dátovú vizualizáciu nie sú zmeny v dátovom priestore nijak zásadné. Koniec-koncov, dáta sú pre vizualizáciu len vstupom a algoritmy by mali spoľahlivo fungovať na všetkých prípustných vstupoch.



Obr. 1. Vizualizačná pipeline [2]

V skutočnosti však majú spomínané revolučné zmeny významný dopad na všetky etapy procesu vizualizácie (Obr.1).

Obrovský objem spracovávaných dát spomaľuje proces filtrovania dát používateľom a rendering, čím trpí hlavne interakcia s človekom.

Integrácia dát z heterogénnych zdrojov (napr. spájanie numerických, textových, vzťahových, časových a geografických údajov) predstavuje veľkú výzvu v procese mapovania, čiže pri prenose údajov do ich grafickej podoby: na pozície, tvary, farby...

Nutnosť zohľadňovať nové atribúty dát, akými sú pravdepodobnosť či dôveryhodnosť, vytvára nové požiadavky na mapovanie a rendering [5].

A nesmieme zabúdať na najdôležitejší element v celom procese: na človeka, ktorý stojí na konci vizualizačnej pipeline ako prijímateľ obrazového výstupu a ktorý prostredníctvom interakcie môže vstupovať do všetkých etáp celého procesu. Kognitívne a perceptuálne obmedzenia ľudského organizmu zostávajú rovnaké bez ohľadu na napredovanie technológií.

3 Inovácie pre vizualizáciu

Spomínané výzvy sú v doméne vizualizácie informácií vo väčšine prípadov už známe a jednotlivé problémy boli skúmané už dávnejšie. Napríklad zobrazovanie heterogénnych dát pomocou prepojených zobrazení [4], rendering dát zohľadňujúci pravdepodobnosť [1] či veľké dáta [6] sú témy, ktorým sa výskum v oblasti vizualizácie venuje už dlho. No až nárast počtu praktických aplikácií, nárast objemu reálnych dát a výskyt viacerých doteraz separátnych problémov súčasne v jednej úlohe preveria navrhnuté postupy v aktuálnych podmienkach.

Atraktívne bude sledovať nový výskum hneď v niekoľkých oblastiach:

3.1 Kolaboratívna vizuálna analýza

Využitie viacerých analytikov či expertov z rôznych oblastí je sľubným riešením problémov spôsobených nárastom objemu a variability analyzovaných dát. Zdieľanie poznatkov medzi účastníkmi kolaboratívnej analýzy predstavuje celkom nové prostredie pre využitie vizualizácie.

3.2 Metadáta

Informácie o spoľahlivosti zdroja dát, dôveryhodnosti, dátume získania ale aj mnohé iné metadáta sú dôležité pri posudzovaní dát a dopyt po zohľadnení metadát vo vizualizácii bude rásť.

3.3 Focus+Context

Rozdelenie dát na focus a kontext je známy a dobre osvedčený postup používaný pre prehľadné zobrazenie veľkých alebo zložitých dát [8]. Bude zaujímavé sledovať jeho aplikáciu na heterogénne dáta, kedy sa focus od kontextu nemusí líšiť len inou úrovňou detailu, ale aj iným zdrojom dát či úplne inými typom mapovania dát.

3.4 Využitie umelej inteligencie

Rozmiestňovanie vrcholov grafu, usporiadanie osí v diagrame, výber mapovaných dátových atribútov a iné parametre vizualizácie majú veľa možných konfigurácií [9]. Algoritmy umelej inteligencie sa od ľudských používateľov môžu učiť, ktoré konfigurácie sú vhodnejšie od iných a uľahčiť tak analytikom ďalšiu prácu.

4 Načo to všetko?

Tieto a mnohé ďalšie výzvy a inovácie budú sprevádzať oblasť vizualizácie informácií v období po nedávnej zmene dátovej paradigmy. Dáta prichádzajú z nekontrolovateľných zdrojov, v bezprecedentnom objeme a často treba reagovať rýchlo. Z oblasti dát a znalostí tak čelíme niekoľkým zásadným výzvam, ktoré zároveň tvoria motiváciu pre ďalší výskum a vývoj v oblasti vizualizácie informácií:

1. Technická výzva: Nárast objemu spracovaných dát, integrácia dát z rôznych zdrojov
2. Technologická výzva: Potreba kolaborácie a integrácie v analytickom procese
3. Politická výzva: Narábanie s dezinformáciami a dátovo-orientovaná žurnalistika.

Nárast objemu spracovaných dát sa technicky darí zvládnuť zlepšeniami hardvéru a data mining algoritmov. No ako povedal John Stasko: „*Data mining použite, keď poznáte otázku. Použite vizualizáciu, keď otázku nepoznate*“ [7]

Kolaboratívny a integračný proces prebieha v komunikácii medzi rôznymi expertmi a analytikmi riešiacimi. Zatiaľ nikto nevymyslel efektívnejší spôsob komplexnej informaçnej výmeny medzi ľuďmi než je obraz.

Nárastu propagandy v médiách treba čeliť pomocou kvalitnej dátovo-orientovanej žurnalistiky. A tá sa nezaobíde bez infografiky a dátovej vizualizácie.

5 Záver

Vizualizácia informácií zostáva dôležitým nástrojom pri objavovaní netušeného, overovaní tušeného a komunikovaní overeného. Vývoj v oblasti získavania, spracovania a sprístupňovania dát tak zákonite vytvára nové výzvy aj pre vizualizáciu. V príspevku sme predstavili niektoré z týchto výziev ako aj viaceré atraktívne výskumné smery, ktoré na ne reagujú.

Literatúra

1. Brodlie, K., Allendes Osorio, R., Lopes, A.: A Review of Uncertainty in Data Visualization. In: *Expanding the Frontiers of Visual Analytics and Visualization*, Springer London (2012), pp. 81-109, 2012.
2. Card, S., Mackinlay, J., Shneiderman, B. (Eds.): *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc (1999)
3. Doleisch, H., Gasser, M., Hauser, H.: Interactive feature specification for focus+context visualization of complex simulation data. In: *Proceedings of the symposium on Data visualization 2003 (VISSYM '03)*. Eurographics Association, pp. 239-248.
4. Kehrer, J., Hauser, H.: Visualization and Visual Analysis of Multifaceted Scientific Data: A Survey, *IEEE Transactions on Visualization and Computer Graphics* (2013), vol. 19, no. 3, pp. 495-513
5. Liu, S., Cui, W., Wu, Y., Liu, M.: A survey on information visualization: recent advances and challenges. *The Visual Computer: International Journal of Computer Graphics* (2014), vol. 30, no. 12, pp. 1373-1393
6. Novotný, M.: *Information Visualization of Large Data*. Faculty of Mathematics, Physics and Informatics, Comenius University Bratislava (2008)
7. Stasko, J.: The Value of Visualization...and Why Interaction Matters. Eurovis 2014 capstone
8. Tominski, Ch., Gladisch, S., Kister, U., Dachsel, R., Schumann, H.: A Survey on Interactive Lenses in Visualization. In: *EuroVis – State-of-the-Art Reports*. The Eurographics Association (2014)
9. Wongsuphasawat, K., Moritz, D., Anand, A., Mackinlay, J., Howe, B., Heer, J.: Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations. In: *IEEE Transactions on Visualization and Computer Graphics* (2016), vol. 22, no. 1, pp. 649-658

Annotation:

Information visualization – current challenges and trends

The increase in data sources and data volume introduces new challenges for information visualization: integration of heterogeneous data sources, large data, uncertainty in data, real-time streaming data etc. While the visualization domain is already familiar with many potential solutions, e.g. linked views, focus+context visualization, or density-based visualization, the abrupt changes in data domain will put these techniques to the real test. And the need for some new techniques arises, such as integration of artificial intelligence into visual analytics and collaborative analysis among multiple human experts.

Graph Mining: Applications

Karel Vaculík^{1,2}

¹Faculty of Informatics Masaryk University
Botanická 68A, 602 00 Brno, Czech Republic

²Gauss Algorithmic s.r.o.
Slovákova 11, 602 00 Brno, Czech Republic

xvaculi4@fi.muni.cz

Abstract. Traditional data mining algorithms typically assume data instances to be independent. However, there is a lot of real-world scenarios where relationships between data instances exist and they are principal for data understanding. For example, there are relationships between people in social networks, between chemical elements in chemical compounds, etc. It is difficult or even impossible to express such information in the classical attribute-value representation. Graph mining is an area of data mining that uses a graph representation of data and it allows us to exploit the relationships in the data. The goal of this talk is to present diverse successful applications of graph mining on real-world graphs.

Talk type: Invited talk

Keywords: graph mining, network analysis, data mining, classification, anomaly detection, community detection, recommendation

1 Introduction

Graph mining is an area of data mining in which data is presented as a graph or a set of graphs [2]. Compared to regular attribute-value representation, graphs allow us to model dependencies between individual entities. These graphs can be either static or dynamic, i.e. they change through time. Nodes and edges can also have attributes assigned. In this paper, we present several applications of graph mining on real-world graphs.

2 Classification of Nodes

The first application we would like to present is node classification in graphs. It is assumed that particular nodes have a class assigned and the task is to train a model for class prediction of other nodes. These models typically utilize the network structure and classify nodes by using their neighbour nodes or the nodes that are similar with regard to a defined measure. Structural neighbourhood-based classifier (SNBC) algorithm [3]

generalizes this basic scenario by allowing multiple classes for each node. By using random-walk technique, the algorithm was able to classify scientific publications on the basis of citation network, categorization of books on the basis co-purchasing network, etc.

3 Anomaly Detection in Recommendation Networks

The next application is treated as a classification task as well, but now the edges are classified [4]. More precisely, the goal is to decide which edges are anomalous and which are not in order to improve the performance of a geospatial recommender, Google Related Places Graph. This recommender system recommends similar or close places (businesses, sights, etc.) for a given place searched via Google Search Engine¹. It uses a network in which *similar* places are linked together. By detecting and removing anomalous edges, the authors were able to filter out plenty of irrelevant recommendations. Anomaly detection is carried out by Random Forests classifier [1] that uses various structural features extracted from the network as well as features from Google Knowledge Graph [6].

4 Anomaly Detection in Communication Networks

Detection of anomalous patterns in dynamic networks is presented in [7]. Patterns are represented by subgraphs that change into other subgraphs in the next moment. By a change, we mean addition or deletion of vertices or edges, change of labels, or a combination of these elemental changes. Thus, the patterns express the changes on the local level of the network.

An example of two patterns from an email-correspondence network is shown in Fig. 1. The nodes represent employees (Emp = regular employee, VP = vice president) and the links represent sent emails. The left part of the patterns depicts the communication on one day and the right part the next day. Frequently occurring patterns are marked as normal whereas deviations from these patterns are marked as anomalies. More specifically, the deviations occur only in the right part of the pattern. The normal patterns capture the common evolution of the subgraphs and they also serve as an explanation of the anomaly deviation. Besides the analysis of email communication, the method was also used to analyse the graphs of resolution proofs created by computer-science students.

¹ <http://www.google.com>.

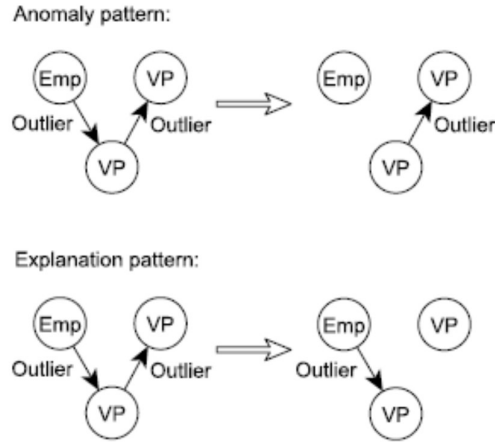


Fig. 1. Anomalous communication pattern and its explanation.

5 Community Detection in Voice-Call Networks

The last application of graph mining is concerned with community detection in a network built from voice calls. Community detection is a process of node clustering in which nodes clustered together are densely connected in the original graph. This work was created during a project in Gauss Algorithmic s.r.o. for a telecommunication company. The nodes in this network represented phone numbers and the edges represented calls between the numbers. More precisely, the edges were obtained by aggregating calls over a longer period of time and weighted by call duration statistics. Label Propagation Algorithm [5] modified for weighted graphs was used for community detection. Subgraphs formed by discovered communities came with various sizes and shapes. Resulting communities are going to be used for improving customer experience and for churn prediction.

6 Summary

In this work we presented four different applications of graph mining on real-world datasets. This is merely a tiny fraction of graph mining. There are many other tasks in this area, such as frequent pattern mining, graph modelling, graph summarization, or link prediction.

Acknowledgements: We would like to thank the organizers for the invitation, Luboš Popelínský and Gauss Algorithmic s.r.o. for support.

Bibliography

1. Breiman, L.: Random Forests. *Machine Learning*, 45(1), 5-32, (2001).
2. Cook, D. J., Holder, L. B.: *Mining Graph Data*. John Wiley & Sons, (2006).
3. Nandanwar, S., Murty, M. N.: Structural Neighborhood Based Classification of Nodes in a Network. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco, California, USA, (2016), pp. 1085–1094.
4. Perozzi, B., Schueppert, M., Saalweachter, J.: When Recommendation Goes Wrong - Anomalous Link Discovery in Recommendation Networks. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco, California, USA, (2016), pp. 569– 578.
5. Raghavan, U. N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. *Phys Rev E* 76:036106, (2007).
6. Singhal, A.: Introducing the knowledge graph: things, not strings. <http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html>, (2012).
7. Vaculik, K., Popelínský, L.: DGRMiner: Anomaly Detection and Explanation in Dynamic Graphs. *Advances in Intelligent Data Analysis XV: 15th International Symposium, IDA*, (2015), (accepted).

Aktuálne dianie v oblasti otvorených údajov v SR (2016)

Peter Hanečák¹, Ľubor Illek²

¹OZ Utopia
Lipského 2, 84101 Bratislava-Dúbravka

²Slovensko.Digital
Staré grundy 12, 841 04 Bratislava

hanecak@opendata.sk, lubor.illek@slovensko.digital

Abstrakt. Téma otvorených údajov v SR zaznamenáva zmiešané reakcie a výsledky, vo všeobecnosti však napreduje a z témy ktorá bola okrajovou až neznámou sa za pár rokov stala téma bežne akceptovaná. Pokiaľ ide o reálne publikovanie, čaká SR ešte mnoho práce, zaznamenali sme však už aj vcelku unikátne a pozitívne výsledky, ktoré predstavujú dobrý základ na ďalší progres.

V prednáške teda bude zhrnutá história otvorených údajov v SR za ostatných zhruba päť rokov (s dôrazom na ostatný rok), aktuálna situácia a tiež odhad toho, čo by sa mohlo udiť v najbližšom období.

Typ príspevku: Pozvaná prednáška

Kľúčové slová: otvorené údaje, Open Data, Slovensko

1 Úvod

Téma otvorených údajov (Open Data) v SR zaznamenáva zmiešané reakcie a výsledky, vo všeobecnosti však napreduje: Po pripojení sa SR k OGP a stanovení si úloh v prvom Akčnom pláne najmä v rámci témy otvorených údajov sa z témy, ktorá bola okrajovou až neznámou, za pár rokov stala téma bežne akceptovaná. Pokiaľ však ide o reálne publikovanie, čaká SR ešte mnoho práce (najmä v oblasti publikovania tzv. prioritných datasetov a dodržiavanie etablovaných štandardov), zaznamenali sme však už aj vcelku unikátne a pozitívne výsledky, ktoré predstavujú dobrý základ na ďalší progres.

V prednáške teda bude zhrnutá história otvorených údajov v SR za ostatných zhruba päť rokov (s dôrazom na ostatný rok), aktuálna situácia a tiež odhad toho, čo by sa mohlo udiť v najbližšom období.

Obsah prednášky sa opiera najmä o pamäť autorov a nepredstavuje kompletný ani vyčerpávajúci zoznam noviniek v téme otvorených údajov. Poradie sekcií je zhruba chronologicky, zohľadňujúc však nie len formálny začiatok aktivít ale to, kedy boli pozorované najviditeľnejšie zmeny

2 Uplynulých zhruba 5 rokov

2.1 Sformovanie komunity OpenData.sk

Komunita OpenData.sk vznikla ako neformálna iniciatíva pod záštitou OZ Utopia v roku 2010. Základom pre iniciatívu boli predchádzajúce aktivity ďalších NGO: SOIT, Aiancia Fair-Play či Transparency International.

2.2 Iniciatíva pre otvorené vládnutie (OGP)

V septembri 2011 sa SR podpisom vtedajšej premiérky Ivety Radičovej pripojilo k Iniciatíve pre otvorené vládnutie (Open Government Partnership – OGP).

Po pripojení SR k OGP bol prijatý Akčný plán na roky 2012-2013¹ (formou uznesenia vlády č. 50/2012²) v ktorom si SR uložila 22 úloh, z čoho 12 sa týkalo prístupu k informáciám (a teda aj problematiky otvorených údajov). Opierajúc sa o tento akčný plán a tiež ďalšie existujúce zákony (zákon o slobodnom prístupe k informáciám – 211/2000 Z.z. a zákon o informačných systémoch verejnej správy – 275/2006 Z.z.) sa v SR formálne započali aktivity štátnej správy v oblasti otvorených údajov. V roku 2012 bol napr. spustený dátový katalóg <http://data.gov.sk> a pomocou neho neskôr zverejnených a zdokumentovaných prvých 161 datasetov (k augustu 2013³). Podrobnosti o plnení plánu možno nájsť napr. v dokumente „Nezávislý hodnotiaci mechanizmus: Slovensko: Hodnotiaca správa 2012- 2013“⁴.

Aktivity pokračovali prijatím a následne plnením druhého plánu na rok 2015⁵. Jedným z dôležitých výsledkov tohto plánu je napr. prieskum ohľadom prioritných datasetov⁶, na základe ktorého boli ako prioritné datasety vyhodnotené tieto:

1. Kataster nehnuteľností (ÚGKK)
2. Výsledky volieb (Štatistický úrad)
3. Údaje zo sčítania obyvateľov, obyvateľov, domov a bytov (Štatistický úrad)
4. Obchodný register (Ministerstvo spravodlivosti)
5. Register adries (Ministerstvo vnútra)
6. Živnostenský register (Ministerstvo vnútra)
7. Dáta o dopravných nehodách (Ministerstvo vnútra, PPZ)
8. Dáta o kriminalite (Ministerstvo vnútra, PPZ)
9. Cestovné poriadky (Ministerstvo dopravy, výstavby a regionálneho rozvoja)

¹ <http://www.otvorenavlada.gov.sk/finalna-verzia-akcneho-planu/>

² <http://www.rokovania.sk/File.aspx/ViewDocumentHtml/Uznesenie-12358?listName=Uznesenie&prefixFile=u>

³ <http://www.otvorenavlada.gov.sk/hodnotenie-iniciativy-pre-otvorene-vladnutie/>

⁴ http://www.opengovpartnership.org/sites/default/files/Slovakia_final_2012_0.pdf

⁵ <http://www.otvorenavlada.gov.sk/akcny-plan-na-rok-2015/>

⁶ <https://github.com/otvorenavlada/akcnyplan2015/tree/master/uloha-03>

10. Poštové smerovacie čísla (Ministerstvo dopravy, výstavby a regionálneho rozvoja, Slovenská pošta)
11. Aktuálny stav a znečistenie životného prostredia (Ministerstvo životného prostredia, SHMÚ)

2.3 Výnos 55/2014

Novela Výnosu o štandardoch pre ISVS [č. 55/2014 Z. z.]⁷ účinná od 15.3.2014 zavádza nový pojem „otvorené údaje“ a definuje k nim aj základné štandardy: formáty CSV a JSON, protokol REST, náležitosti ohľadom licencovania, kvality, atď. ako aj povinnosť zaevidovania datasetov na data.gov.sk.

Keďže dodržiavanie Výnosu je pre verejnú správu uložené zákonom č. 275/2006 Z.z., tak táto novelizácia umožňuje verejnej správe zverejňovať otvorené informácie plne v súlade s platnou legislatívou. Zmeny vo Výnose zároveň reagujú na nedostatky, ktoré boli identifikované pri plnení prvého akčného plánu OGP.

2.4 Register ÚZ

V roku 2014 spustilo Ministerstvo financií Slovenskej republiky novú verziu portálu „Register účtovných závierok“ ktorého súčasťou je aj verejné API⁸. Pomocou portálu a API môžu občania a firmy pristupovať k účtovným informáciám slovenských organizácií. Služba je podľa autora prezentácie unikátom, keďže:

1. v čase spustenia existovalo vo svete zrejme len jedno ďalšie podobné riešenie a to vo Veľkej Británii, pričom slovenské riešenie poskytuje omnoho podrobnejšie informácie než to britské,
2. bolo to zrejme prvé oficiálne Open Data API v SR,
3. otvorené údaje z tohto portálu si našli veľmi rýchlo využitie nielen v slovenskej neziskovej sfére (čo je tradičné) ale aj v podnikateľskej sfére (to až také obvyklé nie je): vznikol portál FinStat.sk.

Zároveň API Registra ÚZ započalo éru užšej spolupráce medzi štátnou správou a Open Data komunitou (viď prezentácie o API a FinStat.SK na stretávke komunity⁹).

⁷ <http://www.informatizacia.sk/standardy-is-vs/596s>

⁸ <http://www.registeruz.sk/cruz-public/version/193084/static/api.html>

⁹ <https://utopia.sk/wiki/display/opendata/OpenData.sk+Meetup+%233#OpenData.skMeetup%233-BlokoAPIRegistraÚZ>

3 Ostaný rok

3.1 DanubeHack

V dňoch 15. až 17.10.2015 sa v Bratislave konal Danube Open (Geo) Data Hackathon & Developers' Workshops, v skratke známy ako DanubeHack¹⁰. Bol to prvý hackaton v SR ktorý sa venoval otvoreným údajom a zároveň bol spoluorganizovaný štátnou správou, konkrétne Slovenskou agentúrou životného prostredia (SAŽP) a Národnou agentúrou pre sieťové a elektronické služby (NASES), ktoré do hackatonu pripravili aj nové otvorené údaje (SAŽP údaje z domény GEO a NASES údaje z Registra adries).

Hackaton mal medzinárodnú účasť, súťažilo na ňom 9 projektov a víťazi získali ceny v hodnote 3000€. Podrobnejšie informácie o výsledkoch možno nájsť na stránkach hackatonu¹¹.

3.2 Nová verzia data.gov.sk

Koncom roka 2015 resp. začiatkom roka 2016 bola do prevádzky spustená nová verzia portálu data.gov.sk. Okrem viditeľných zmien (nová grafika, novšia verzia softvéru CKAN, atď.) v sebe tento upgrade pod hlavičkou projektu eDemokracia/Modul Open Data (MOD) prináša aj komplexnú Open Data infraštruktúru pre štátnu ale aj verejnú správu, integrovanú s existujúcim Ústredným portálom verejnej správy (ÚPVS, <http://slovensko.sk>). V novej verzii totiž MOD obsahuje aj nástroje na ukladanie a publikovanie otvorených údajov priamo na portály data.gov.sk, transformačné a vizualizačné nástroje ako aj ucelenejšiu podporu procesov publikovania otvorených údajov¹².

3.3 Zvýšená aktivita Štatistického úradu

V roku 2015 Štatistický úrad (ŠÚ) výrazne zvýšil svoje aktivity ohľadom zverejňovania otvorených údajov. Keďže poskytovanie informácií verejnosti je vlastne jednou z hlavných úloh úradu, tak ťažiskom ich aktivít ohľadom otvorených údajov bola a je konverzia údajov z týchto systémov do otvorených formátov v súlade s Výnosom 55/2014, t.j. automatizovanie exportu DATAcube kociek do formátu CSV, voľba licencie CC-BY-SA a evidovanie datasetov na portály data.gov.sk¹³.

K 7.9.2016 má na data.gov.sk ŠÚ zaevidovaných 606 datasetov z celkového počtu 1002. Aktuálne informácie možno sledovať priamo na data.gov.sk¹⁴.

¹⁰ <http://www.danubehack.eu/>

¹¹ <http://www.danubehack.eu/?#section-results>

¹² <https://www.nases.gov.sk/data/files/9218.pdf>

¹³ <https://utopia.sk/wiki/display/opendata/OpenData.sk+Meetup+%236>

¹⁴ <https://data.gov.sk/organization/f4787c6f-9fa3-406c-b8d5-d374f1e1f2d3>

3.4 Zapojenie samosprávy

V roku 2015 zavŕšilo mesto Prešov svoje viac než štvorročné Open Data aktivity spustením katalógu otvorených údajov¹⁵ a zaevidovaním svojich datasetov na data.gov.sk¹⁶. Následne svoje znalosti a skúsenosti začali zdieľať aj s ďalšími mestami a tak napr. počiatkom tohto roku (2016) spustilo svoj katalóg otvorených údajov ako aj zverejňovanie datasetov mesto Levice¹⁷.

Do roku 2015 boli otvorené údaje doménou štátnej správy pričom samospráva o možnostiach a povinnostiach ohľadom otvorených údajov dá sa povedať netušila. Od roku 2015 môžeme ohľadom otvorených údajov začať hovoriť aj o aktívnom zapojení slovenských samospráv.

3.5 Register Adries ako Open Data

Začiatkom roka 2016 bolo do oficiálnej prevádzky spustené publikovanie údajov z Registeru Adries (RA) vo forme otvorených údajov prostredníctvom portálu data.gov.sk¹⁸. Je to jeden z prvých pilotných projektov publikovania údajov prostredníctvom už spomínaného Modulu Open Data (MOD) v rámci ktorého Ministerstvo vnútra SR na základe dohody s NASES poskytuje údaje z RA prostredníctvom „interného“ G2G API do data.gov.sk pričom práve data.gov.sk vykonáva formátové a obsahové konverzie potrebné na to, aby údaje spĺňali náležitosti Výnosu 55/2015 ohľadom otvorených údajov.

Zverejňovanie údajov z RA je jedným z modelov spolupráce, ktorý ponúka NASES ostatným organizáciám verejnej správy, kedy tieto organizácie poskytnú údaje v nezmenenej podobe a NASES zabezpečí riadne zverejnenie vo forme otvorených údajov v súlade s platnými štandardami, čím sa zabezpečuje správne a zároveň efektívne údaje.

Poznámka: Existuje viacero správnych spôsobov zverejňovania otvorených údajov a tak NASES ponúka aj viacero modelov spolupráce pri zverejňovaní. Na jednej strane je možné zverejňovanie čisto v réžii tzv. povinných osôb (PO), kedy NASES aktívne nespolupracuje a poskytuje iba minimálne služby potrebné na evidovanie datasetov na data.gov.sk (vyžadované Výnosom 55/2014). Na opačnej škále spolupráce je vyššie popísaný model, kedy PO robí len minimálne úkony a drvivú väčšinu práce vykoná NASES.

3.6 Referencovateľný identifikátor a Linked Data vo Výnose 55/2014

Od roku 2015 prebieha proces aktualizácie Výnosu 55/2014 ohľadom Linked Data a referencovateľných identifikátorov. Stavia sa na novele z roku 2014 a cieľom je doplniť existujúci štandard tzv. dátových prvkov vo formáte XML (príloha č. 2 k Výnosu) aj o

¹⁵ <https://utopia.sk/wiki/pages/viewpage.action?pageId=58360521>

¹⁶ <https://data.gov.sk/organization/8a043e04-3ef8-45d4-a9a9-ede214e5fac5>

¹⁷ <https://utopia.sk/wiki/pages/viewpage.action?pageId=61866089>

¹⁸ <https://data.gov.sk/dataset?tags=register+adries>

reprezentáciu v RDF pomocou zavedenia slovenskej ontológie pre dátové prvky a ich napojenie na vo svete zaužívané ontológie. Špecifická slovenská ontológia je motivovaná tým, aby bol prechod z XML na RDF jednoduchý (napr. pomocou jednoduchých XSLT transformácií). Zároveň však budú využité vlastnosti Linked Data a Semantic Web tak, aby boli slovenské údaje prepojené a prepojitelné na údaje v zahraničí.

Aktuálne je návrh novely prejednávaný v rámci PS1 Štandardizačnej komisie¹⁹ a tiež aj v širšej komunite²⁰.

3.7 Akčný plán OGP 2017-2019

Úrad splnomocnenca vlády pre rozvoj občianskej spoločnosti (ÚSV ROS) v spolupráci s občanmi pripravil ďalší akčný plán OGP pre roky 2016 až 2019²¹. Akčný plán definuje 69 úloh z ktorých 14 v kategóriách otvorené údaje a otvorené API.

Úlohy v kategórii otvorených údajov predstavujú kontinuitu k predchádzajúcim plánom a možno ich vnímať ako inkrementálne zlepšovanie zverejňovania otvorených údajov v SR pričom dôraz sa presúva od technických otázok k zlepšovaniu procesov a kvality.

Úlohy spojené s otvorenými API predstavujú revolučný krok vpred, vďaka ktorému by sa mal odomknúť skrytý potenciál základných elektronických služieb verejnej správy tým, že k nim dostanú možnosť pristupovať aj firmy, neziskové organizácie alebo jednotlivci vďaka čomu budú schopní občanom ponúknuť výrazne rozšírené služby, ktoré pomôžu občanom aj v situáciách, ktoré plne nespádajú do pôsobnosti verejných inštitúcií.

Príkladom môže byť kúpa auta: Dnes si pomocou elektronických služieb štátu môže občan cez internet vybaviť základné úradné úkony (prihlásenie vozidla do evidencie, atď.) ale rôzne ďalšie povinnosti (napr. povinné zmluvné poistenie) alebo doplnkové veci (napr. havarijné poistenie) si musí vybaviť inde. V tomto prípade môže rozšírená služba ponúknutá firmou ponúknuť občanovi všetko na jednom mieste a výrazne jednoduchšie (od prihlásenia vozidla až po havarijné poistenie) – štát niečo také poskytnúť nemôže (nemá v kompetencii ponúkať komerčné havarijné poistenie) a vďaka elektronickému občianskemu preukazu (eID) a zaručenému elektronickému podpisu (ZEP) môže byť riešenie stále plne bezpečné aj keď bude prostredníkom súkromná firma.

3.8 Iniciatíva Slovensko.Digital

Koncom roka 2015 Slovensko dosiahlo jeden míľnik v informatizácii verejnej správy: z prostriedkov EÚ ako aj vlastných prostriedkov SR bolo od roku 2007 vynaložených viac ako 900 miliónov € na rôzne informačné systémy, z pohľadu občana však vidno len málo prínosných riešení. Takáto bilancia dala dokopy niekoľko stoviek nadšencov

¹⁹ <https://wiki.finance.gov.sk/label/PS1/ps1-19>

²⁰ <https://platforma.slovensko.digital/t/semanticke-datove-standardy-pre-udaje-verejnej-spravy-sr/185>

²¹ http://www.minv.sk/?ros_ogp_tvorb_2016-19&sprava=finalny-navrh-akcneho-planu-iniciativy-pre-otvorene-vladnutie-na-roky-2016-2019-zverejneny

z IT sektora ale aj verejnej správy a sformovali platformu Slovensko.Digital, ktoré sa začiatkom roka transformovalo na oficiálnu neziskovú organizáciu ktorá si stanovila za cieľ spolupracovať s verejnými inštitúciami na zvýšení kvality ich digitálnych služieb²².

Jedným zo spôsobov ako to dosiahnuť je ukazovať pozitívne príklady, ako sa dá robiť štátne IT dobre. Jednou z inšpirácií je napr. Ekosystém.Slovensko.Digital²³, ktorý bol uverený do testovacej prevádzky dňa 18.8.2016²⁴. Ekosystém ukazuje, že je možné prioritné otvorené údaje (Register právnických osôb, Centrálny register zmlúv a ďalšie) publikovať jednoducho, lacno a dobre (pomocou CSV a BitTorrent) a zároveň k nim jednoducho, lacno a dobre poskytovať aj API (na báze REST A SQL) a ďalšie doplnkové služby (Autoform, ktorý poskytuje automatické dopĺňanie informácií o firmách pri vyplňaní formulárov práve na základe oficiálnych otvorených údajov).

4 Aktuálna situácia

V roku 2016 nastal na oficiálnej úrovni možno v otázkach otvorených údajov konštatovať útlm zrejme spôsobený najprv očakávaním volieb, neskôr výsledkom volieb a ešte neskôr predsedníctvom SR v Rade EÚ.

Prebieha napríklad presúvanie kompetencií ohľadom informačných systémov verejnej správy (ISVS) na novovytvorený Úrad podpredsedu vlády SR pre investície a informatizáciu ktorý vedie p. Peter Pellegrini^{25 26}.

V príprave je nový akčného plánu OGP pre roky 2016 až 2019 (spomínaný vyššie).

A existuje aj návrh „Stratégia a akčný plán sprístupnenia a používania otvorených údajov verejnej správy“ vypracovaný v NASES²⁷, ktorý napr. medzi základné ciele navrhuje naďalej sa pridržať základného princípu „zverejňovanie a sprístupňovanie všetkých dát verejnej správy, ktoré nie sú utajené alebo chránené, a „publikovanie štruktúrovaných datasetov“. V rámci konkrétnych opatrení sú navrhované doplnenia štandardov ISVS v časti otvorených údajov, zriadenie role dátového kurátora a medzirezortný podporný tím pre zverejňovanie údajov.

5 Výhľad do budúcnosti

Ako občania a aktivisti môžeme v otázkach zverejňovania otvorených údajov verejnými inštitúciami poskytnúť len odborný odhad: Odhadujeme, že novovytvorený Úrad podpredsedu vlády SR pre investície a informatizáciu v spolupráci s občanmi (Sloven-

²² <https://slovensko.digital/>

²³ <https://ekosystem.slovensko.digital/>

²⁴ <https://platforma.slovensko.digital/t/ekosystem-slovensko-digital-novinky/2326>

²⁵ <https://www.vicempremier.gov.sk/index.php/informatizacia/index.html>

²⁶ <https://www.vicempremier.gov.sk/index.php/o-urade/organizacna-struktura/index.html>

²⁷ https://www.nases.gov.sk/data/files/20160122_OpenData_vpk.pdf

sko.Digital a ďalšími) si ako jednu z logických priorít v nasledujúcom období pri rozvoji informačných systémov verejnej správy ako jednu z priorít zvolí otvorené údaje a že nastane posun v dôležitých otázkach:

- publikovanie prioritných datasetov na základe verejných konzultácií (úloha plynúca z návrhu akčného plánu OGP)
- zlepšovanie štandardov zverejňovania datasetov (Výnos 55/2014)
- vynucovanie dodržiavania štandardov v existujúcich ale najmä nových informačných systémoch
- aktivity smerujúce k reálnemu ekonomickému zhodnoteniu zverejňovaných otvorených údajov (úspory alebo nová pridaná hodnota), keďže to je jedna z motivácií verejnej správy („Prínos do rozvoja ekonomiky štátu je jedným z hlavných motívov EÚ pre sprístupňovania Open Data.“²⁸) ale aj motivácia komunity či firiem.

6 Záver

Téma otvorených údajov má v SR už svoju históriu a aj zaujímavé výsledky. Aktuálna utlmená aktivita verejnej správy je vyvážená novými aktivitami občanov a vďaka doterajším výsledkom a aj novým aktivitám môžeme v budúcnosti očakávať ďalšie pozitívne výsledky.

Podakovanie: Ďakujeme všetkým, ktorý k publikovaniu otvorených údajov v SR dopomáhali v minulosti, či už v rámci občianskych iniciatív alebo „zvnútra“ verejnej správy.

Annotation:

Current Open Data activities in Slovak Republic

Open Data theme is received with mixed reactions and produced varying results. But in general, work related to Open Data is progressing forward and the topic, which was on the fringes of interest, become in few years commonly accepted. In terms of actual publication of Open Data, Slovakia has still long road ahead but noteworthy, even unique results were already achieved. And those achievements form a good base for good progress also in the future.

Presentation contains summary of history of Open Data in Slovakia for past roughly 5 years (with main focus on activities in past year), view on current situation and estimate of what might be happening in near future.

²⁸ https://www.nases.gov.sk/data/files/20160122_OpenData_vpk.pdf

Aktuální dění v oblasti otevřených dat v ČR

Dušan Chlapek¹, Michal Kubáň²

¹Fakulta informatiky a statistiky
Vysoká škola ekonomická v Praze nám. W. Churchilla 4,
130 67 Praha 3, Česká republika

²Ministerstvo vnitra,
náměstí Hrdinů 1634/3, Praha 4, 140, Česká republika

chlapek@vse.cz, michal.kuban@mvcv.cz

Abstrakt. V přednášce budou shrnuty hlavní dosažené výsledky v oblasti otevřených a propojených dat realizované v uplynulých letech v ČR. Dále budou prezentovány záměry dalšího rozvoje oblasti otevřených propojitelných dat v ČR.

Typ příspěvku: Zvaná přednáška

Klíčová slova: otevřená data, Open Data, veřejná správa, Česká republika.

1 Úvod

Otevřenými daty se od 6.9.2016, kdy podepsal novelu zákona č. 106/1999 Sb., o svobodném přístupu k informacím prezident České republiky, rozumí [7] „*informace zveřejňované způsobem umožňujícím dálkový přístup v otevřeném a strojově čitelném formátu, jejichž způsob ani účel následného využití není omezen a které jsou evidovány v národním katalogu otevřených dat*“.

Otevřená data se tak stala, po dvou letech intenzivního úsilí ze strany veřejné správy, zejména Ministerstva vnitra, a ze strany neziskových organizací, součástí právního prostředí i v České republice a mohou být tak využívána jako jeden z důležitých inovativních nástrojů ve fungování veřejné správy. Důležitost otevřených dat je v České republice charakterizována i tím, že se otevřená data dostala mezi cíle v důležitých strategických dokumentech České republiky:

- Strategický rámec rozvoje veřejné správy České republiky pro období 2014-2020 [9], konkrétně jeho Specifický cíl 3.1 – Dobudování funkčního rámce eGovernmentu.
- Akční plán České republiky Partnerství pro otevřené vládnutí na období let 2016 až 2018 [2], část 4.2 - Zpřístupnění dat a informací.

- Strategie rozvoje ICT služeb veřejné správy a její opatření na zefektivnění ICT služeb schválená usnesením vlády č. 889/2015 [10] a její strategický cíl č. 5 - Od izolovaných dat k propojeným a otevřeným datům veřejné správy a ke kvalifikovaným rozhodnutím vedoucím k vyšší efektivnosti služeb VS.
- Akční plán boje s korupcí na rok 2016 [1] a jeho cíl č. 2 - Transparentnost a otevřený přístup k informacím.
- Státní politika v elektronických komunikacích „Digitální Česko v. 2.0, Cesta k digitální ekonomice“ [8] a její cíl 5.4. - Využívání informací veřejného sektoru.
- Akční plán pro rozvoj digitálního trhu [3] a jeho cíl Přístup k datům veřejného sektoru v kapitole 5 : Nové trendy.

Otevřená data se tak stala jednou z oblastí, ve které došlo za poslední dva roky k poměrně významné akceleraci aktivit a výraznému rozšíření povědomí o účelnosti a potřebnosti otevřená data používat jak na úrovni státu, tak i na úrovni samosprávných celků.

2 Aktivita realizované v uplynulých letech

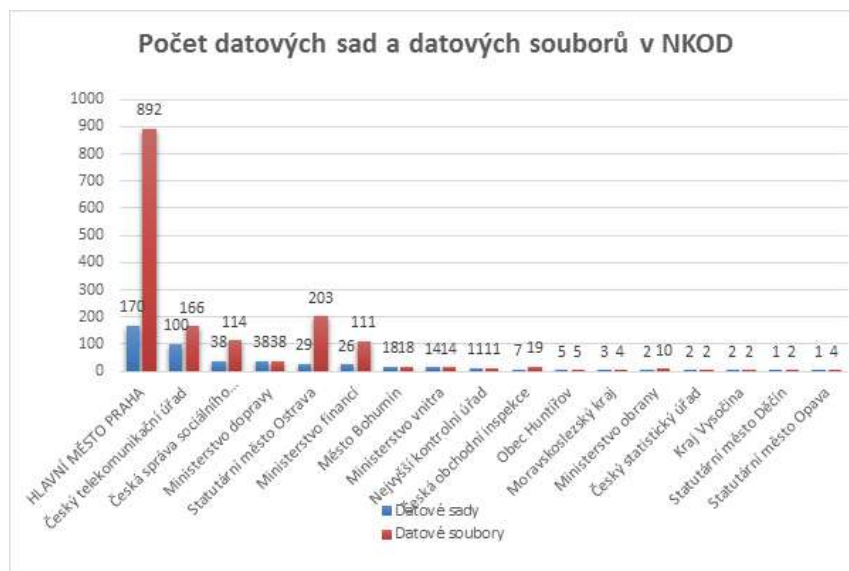
Myšlenka otevřených dat se v ČR začala výrazněji prosazovat v roce 2012, kdy česká vláda schválila Akční plán Partnerství pro otevřené vládnutí, ve kterém se zavázala mimojiné zavázala pro publikaci vybraných informací veřejného sektoru formou otevřených dat. Následně se vláda začala tímto tématem zabývat a vznikla Koncepce katalogizace otevřených dat veřejné správy [5] a Metodika publikace otevřených dat veřejné správy ČR [6]. Nejvýznamnějším milníkem pro rozvoj otevřených dat v ČR byl však rok 2015, kdy se podařilo:

- zahájit plný provoz Národního katalogu otevřených dat (květen 2015) viz <http://data.gov.cz>,
- vytvořit standardy pro přípravu, publikaci a katalogizaci otevřených dat veřejné správy ČR – viz <http://opendata.gov.cz>,
- vytvořit a se zástupci veřejné správy validovat vzorové publikační plány pro jednotlivé typy orgánů veřejné moci, tj. centrální orgány, kraje a jednotlivé typy obcí,
- vytvořit návrh úprav legislativy pro otevřená data VS ČR,
- připravit a realizovat první vlnu vzdělávání v oblasti otevřených dat, vytvořené školicí materiály byly poskytnuty všem úřadům veřejné správy a veřejnosti prostřednictvím webového portálu MV ČR. Celkem bylo proškoleny 415 osob z celkem 206 subjektů, z toho: 10 ministerstev, 7 ostatních ústředních orgánů státní správy, 8 krajů, 69 obcí s rozšířenou působností a 112 ostatních obecních úřadů.

3 Aktuální dění v oblasti otevřených dat

V roce 2016 se práce v oblasti otevřených dat ještě zintenzivnila. Je dokončován legislativní proces, tj. na novelizovaný zákon č. 106/1996 Sb. ještě navazují práce na připravě nařízení vlády, které uloží povinné zveřejnění vyjmenovaných informací jako otevřená data. Nařízení vlády je v současné době (září 2016) v meziresortním připomínkovacím řízení.

V roce 2016 došlo ke zřízení a obsazení pozice Národního koordinátora otevřených dat na Ministerstvu vnitra. Díky obsazení této pozice dochází ke koordinaci a sjednocení metodické podpory pro všechny orgány veřejné správy. Dále je podporován a rozvíjen národní katalog otevřených dat. K začátku září 2016 je v národním katalogu otevřených dat (NKOD) zaevidováno 38963 datových sad, z nich největší počet datových sad má registrován Český úřad zeměměřičský a katastrální (ČÚZK).



Obr. 1. Počty registrovaných datových sad v NKOD bez sad ČÚZK.

Dále jsou v roce 2016 realizovány a připravovány následující aktivity:

- pokračování ve vzdělávání v oblasti otevřených dat a výměně zkušeností,
- rozvoj a veřejné konzultace standardů (workshopy) v oblasti publikace a katalogizace otevřených dat, zejména ve vztahu k vývoji mezinárodních standardů a metodik,
- konzultace specifických datových oblastí a datových sad.

4 Závěr

V dalších letech se Ministerstvo vnitra ve spolupráci s odbornou veřejností a akademickými institucemi zaměří na výraznou inovaci národního katalogu otevřených dat jak po stránce uživatelské přívětivosti, kvality dat tak po stránce interoperability s dalšími katalogy otevřených dat. Dále je zamýšleno vytvoření koncepce zasazení otevřených a propojených dat do Národního architektonického plánu, a to včetně

- vytvoření datové politiky popisující jednotný způsob zasazení principů otevřených dat do kontextu Národního architektonického plánu [4],

- vytvoření koncepce sémantického slovníku pojmů a inicializace jeho tvorby a rozvoje za účelem ujednocování pojmosloví a datových struktur (na syntaktické a především sémantické úrovni) používaných při publikaci otevřených dat,
- vytvoření předpokladů pro propojování otevřených dat jednotlivých orgánů veřejné správy.

Odpovídajícím způsobem budou rozvíjeny Standardy publikace a katalogizace otevřených dat. Při plnění těchto cílů bude kladen důraz na zajištění souladu s aktuálními mezinárodními standardy a metodikami pro oblast otevřených dat, především těch vydávaných Evropskou komisí. Dále bude nastaven proces monitorování stavu plnění koncepce jednotlivými orgány veřejné správy.

Literatura

1. Akční plán boje s korupcí na rok 2016. Dostupné z: <https://www.korupce.cz/assets/protikoru-puci-dokumenty-vlady/na-leta-2015-2017/Akcni-plan-boje-s-korupci-na-rok-2016.pdf>.
2. Akční plán České republiky Partnerství pro otevřené vládnutí na období let 2016 až 2018. Dostupné z: <https://www.korupce.cz/assets/partnerstvi-pro-otevrene-vladnuti/Akcni-plan-Ceske-republiky-Partnerstvi-pro-otevrene-vladnuti-na-obdobi-let-2016-az-2018.pdf>.
3. Akční plán pro rozvoj digitálního trhu. Dostupné z: <http://docplayer.cz/2947416-Urad-vlady-cr-akcni-plan-pro-rozvoj-digitalniho-trhu.html>.
4. Felix, O., Kuchař, P. Národní architektonický plán eGovernmentu ČR. Cíle, stav, budoucnost. Dostupné z: http://www.google.cz/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&ved=0ahUKEwjF18Hi0YD-PAhXEuBQKHYS6CiYQFggnMAI&url=http%3A%2F%2Fwww.cssi.cz%2Fcssi%2Fsystem%2Ffiles%2Fall%2FISI_2015_01_Kuchar_Felix.pdf&usg=AFQjCNGMeozY4pCSnTDl6R1L9E2sNtegRQ.
5. Chlapek, D., Kučera, J., Nečaský, M.: Koncepce katalogizace otevřených dat VS ČR (zkrácená verze), 2012. Dostupné z: <http://www.mvcr.cz/soubor/koncepce-katalogizaceotevrenych-dat-vs-cr-pdf.aspx>
6. Chlapek, D., Kučera, J., Nečaský, M.: Metodika publikace otevřených dat veřejné správy ČR verze 1.0, 2012. Dostupné z: <http://www.mvcr.cz/soubor/metodika-publ-opendata-verze-1-0-pdf.aspx>
7. Sněmovní tisk č. 764/0. Vládní návrh zákona, kterým se mění některé zákony v souvislosti s přijetím zákona o službách vytvářejících důvěru pro elektronické transakce, zákon č. 106/1999 Sb., o svobodném přístupu k informacím, ve znění pozdějších předpisů, a zákon č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů. Dostupné z: <http://www.psp.cz/sqw/tisky.sqw?O=7&T=764>.
8. Státní politika v elektronických komunikacích „Digitální Česko v. 2.0, Cesta k digitální ekonomice“. Dostupné z: <http://www.mpo.cz/dokument127530.html>
9. Strategický rámec rozvoje veřejné správy České republiky pro období 2014-2020. Dostupné z: <http://databaze-strategie.cz/cz/mv/strategie/strategicky-ramec-rozvoje-verejne-spravy-ceske-republiky-pro-obdobi-2014-2020>.
10. Strategie rozvoje ICT služeb veřejné správy a její opatření na zefektivnění ICT služeb. Dostupné z: <https://apps.odok.cz/zvlady/usneseni/-/usn/2015/889>.

Annotation:

Open Data activities in the Czech Republic

The objective of this paper is to summarize the open data activities in the Czech Republic in the recent years. It also highlights the milestones of implementation of open data into Czech legal system as well as the role and use of created open data methodology and standards for publication and catalogisation of open data for Czech state administration. Furthermore, the paper also presents planned activities of the Ministry of the Interior of the Czech Republic in this agenda.

Rozpoznání atributů vozidel pomocí metod strojového učení

Jan Sedlák

Fakulta informatiky, Masarykova univerzita Botanická 68a,
602 00 Brno, Česká republika

`xsedlak5@fi.muni.cz`

Abstrakt. Schopnost detekce a přesného rozpoznání různých atributů, jako jsou například barva nebo výrobce vozidla, hraje poměrně důležitou roli v inteligentních dopravních systémech (ITS), ale i při práci PČR, kde je tato schopnost velmi ceněna, obzvláště při detekci zájmových nebo odcizených vozidel. Práce se zaměřuje na klasifikaci zmíněných atributů ze snímků získaných z různých kamer v reálném provozu. Takovéto snímky často obsahují různé typy deformací, které správnou klasifikaci výrazně komplikují. Součástí práce je praktické porovnání použitelnosti populárních metod strojového učení, mezi které patří například RandomForest, Support vector machine a nebo dnes stále oblíbenější hluboké neuronové sítě. Provedené experimenty ukázaly, že ačkoli hluboké neuronové sítě dosahují velmi dobrých výsledků, ne vždy je nutné a efektivní tuto metodu využít.

Typ příspěvku: Zvaná přednáška

Klíčová slova: strojové učení, neuronové sítě, klasifikace výrobce, barva vozidla, SVM, RandomForest, hluboké neuronové sítě

1 Úvod

Schopnost detekce a přesného rozpoznání různých atributů, jako jsou například barva nebo výrobce vozidla, hraje poměrně důležitou roli v inteligentních dopravních systémech (ITS), ale i při práci PČR, kde je tato schopnost velmi ceněna obzvláště při detekci zájmových nebo odcizených vozidel. Znalost podrobnějších informací o detekovaném vozidle výrazně pomáhá při jeho zpětném vyhledání (například při průjezdu zájmového vozidla), kde tyto informace mohou zpřesnit výsledek hledání. Základním atributem při hledání je v obecném případě registrační značka (dále jen RZ) vozidla.

V případě, že její detekce selže (například vlivem nesplnění vstupních podmínek použitého algoritmu, úmyslného zastínění nebo odstranění RZ apod.), je možné následnou identifikaci zjednodušit použitím ostatních zjištěných atributů o vozidle. Tyto atributy navíc umožňují automatickou detekci událostí, které při použití pouze registrační

značky nejsou možné, jedná se zejména o neoprávněné použití jedné RZ na více vozidlech.

2 Obsah přednášky

V přednášce se zaměřím na klasifikaci barvy a rozpoznání typu vozidla s použitím různých metod strojového učení pro řešení těchto problémů. Datová sada je složena ze snímků, které jsou pořízeny z různých kamer nasazených v reálném provozu. Kvalita takto získaných snímků je velmi různorodá a často je ovlivněna vnějším prostředím, vlivem kterého dochází k různým deformacím pořízeného snímku. Mezi další faktory ovlivňující kvalitu snímku patří i různé natočení kamery nebo například i samotný typ kamery.

Nejčastějším způsobem klasifikace různých atributů vozidla je postupná segmentace obrazu na zájmové oblasti, které jsou následně zpracovány a samostatně klasifikovány. Prvním krokem je zpravidla detekce vozidla, rozpoznání RZ a následně upřesnění pozice vozidla. Znalost pozice a úhlu natočení registrační značky může být využita pro implementaci algoritmu, který efektivně provede výběr zájmové oblasti. Ze získané zájmové oblasti jsou následně extrahovány informace, které jsou poté klasifikovány pomocí metod strojového učení.

V průběhu přednášky bude předvedeno praktické porovnání použitelnosti populárních metod strojového učení, mezi které patří RandomForest, Support vector machine (SVM) a dnes stále populárnější hluboké neuronové sítě. Tyto vybrané metody strojového učení jsou aplikovány na snímky získané z více jak 50-ti kamer.

Na závěr budou shrnuty výhody a nevýhody jednotlivých metod, včetně získaných zkušeností z reálného provozu.

Literatura

1. Reza Fuad Rachmadi a I Purnama. „Vehicle Color Recognition using Convolutional Neural Network“. In: arXiv preprint arXiv:1510.07391 (2015). url: <http://arxiv.org/pdf/1510.07391.pdf>.
2. Chuanping Hu et al. „Vehicle Color Recognition With Spatial Pyramid Deep Learning“. In: Intelligent Transportation Systems, IEEE Transactions on 16.5 (2015), s. 2925–2934
3. Lisa M Brown, Amitava Datta a Sharath Pankanti. „Tree-based vehicle color classification using spatial features on publicly available continuous data“. In: Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on. IEEE. 2013, s. 347–352.
4. Yongbin Gao a Hyo Jong Lee. „Vehicle Make Recognition Based on Convolutional Neural Network“. In: Information Science and Security (ICISS), 2015 2nd International Conference on. IEEE. 2015, s. 1–4.
5. Li-Chih Chen et al. „Vehicle make and model recognition using sparse representation and symmetrical SURFs“. In: Pattern Recognition 48.6 (2015), s. 1979–1998
6. Michael A. Nielsen. Neural Networks and Deep Learning. <http://neuralnetworksanddeeplearning.com/chap5.html>. 2015.

7. Geoffrey E. Hinton et al. „Improving neural networks by preventing co-adaptation of feature detectors“. In: CoRR abs/1207.0580 (2012). url: <http://arxiv.org/abs/1207.0580>.

Annotation:

Automatic recognition of vehicle attributes using machine learning

Detection and recognition of vehicle attributes is important part of the ITS systems, particularly when searching for interest or stolen vehicles. The lecture discusses applicability of machine learning methods, including RandomForest, Support vector machine and deep neural networks, in identifying the individual vehicle attributes based on camera images from the real environment. The results from functional implementation of machine learning algorithms for classification of color and vehicle make, deployed in real-world environment for the purpose of Police of the Czech Republic, will be presented.

Aplikácia prístupov hlbokého učenia na riešenie (ne)štandardných úloh strojového učenia

Michal Barla, Peter Lacko, Mária Šajgalík

Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava, Slovenská republika

{michal.barla, peter.lacko, marius.sajgalik}@stuba.sk

Abstrakt. Veľké medzinárodné spoločnosti ako Google, Microsoft alebo Facebook investujú nemalé prostriedky do podpory výskumu a rozvoja oblasti hlbokých neurónových sietí. Nedávne porazenie šampióna v hre Go, práve vďaka hlbokému učeniu, ukazuje potenciál tohto prístupu. Aplikácie strojového učenia založené na hlbokých umelých neurónových sieťach dosahujú v mnohých oblastiach lepšie výsledky ako prístupy založené na ručne ladených črtách. V tejto pozvanej prednáške si prejdeme základné princípy hlbokého učenia a ukážeme si aplikáciu tohto prístupu na rôzne problémy, ktoré riešime na UISI FIIT STU.

Typ príspevku: Pozvaná prednáška

Kľúčové slová: neurónové siete, strojové učenie

1 Úvod

Popularita umelých neurónových sietí v doméne strojového učenia v poslednom čase významne vzrástla. Je to hlavne vďaka novým úspešne použitým prístupom učenia a architektúram neurónových sietí, ako aj dostupnosťou masívne paralelných výpočtových prostriedkov na tréning sietí (čipov grafických kariet). Zásahu na tom má aj to, že veľké medzinárodné spoločnosti ako Google, Microsoft alebo Facebook investujú nemalé prostriedky do podpory výskumu a rozvoja v tejto oblasti. Tento príspevok opisuje krátke predstavenie hlbokých sietí a aplikácie tohto prístupu na rôzne problémy, ktoré riešime na UISI FIIT STU.

2 Od perceptrónu ku hlbokkej neurónovej sieti

Klasifikácia objektov, ako napríklad rozpoznávanie vzorov, medicínskych diagnóz a pod. je veľmi frekvencovaný problém riešený v doméne strojového učenia. Pri týchto úlohách máme množinu označených objektov (vektorov vlastností, čo môžu byť body obrázku, alebo výsledky testov pacienta). Ku každému objektu máme priradenú aj jeho kategóriu, pes, mačka, alebo chrípka, angína.

Základným stavebným blokom neurónovej siete je neurón a jedným z prvých prístupov učenia bolo použité perceptrónu. Perceptrón je v najjednoduchšej podobe binárny klasifikátor, ktorý mapuje vstupy \mathbf{x} na výstupné hodnoty 0 alebo 1 podľa:

$$h(\mathbf{x}) = \begin{cases} 1 & : \text{ak } \mathbf{w} \cdot \mathbf{x} + b > 0 \\ 0 & : \text{inak} \end{cases},$$

kde \mathbf{w} je vektor váh a b je prahová hodnota. Perceptrón sa trénuje pomocou zmeny váh tak, že pre každý vstupný vektor z trénovacej množiny sa vypočíta výstup perceptrónu a váhy sa upravujú tak, aby sa minimalizovala chyba klasifikácie.

Nevýhodou samostatného perceptrónu, je že je schopný naučiť sa len lineárne separovateľné problémy. Preto napríklad nie je schopný aproximovať funkciu XOR. Riešením tohoto problému je vytvorenie viacvrstvovej doprednej siete, čo je vlastne len pospájanie viacerých perceptrónov do vrstiev. Najjednoduchšou viacvrstvou sieťou je dopredná trojvrstvomá sieť, kde prvá vrstva predstavuje vstup siete, druhá vrstva, ktorá sa nazýva aj skrytou obsahuje perceptróny ktoré spracovávajú aktivácie zo vstupov a výstupná (tretia) vrstva spracováva informácie zo skrytej vrstvy. Prepojenie medzi vrstvami je typicky úplné, perceptrón z vrstvy spracováva aktivácie všetkých neurónov v predchádzajúcej vrstve. Každé prepojenie medzi neurónmi je definované váhou w . Spracovanie vstupných dát prebieha po vrstvách. Na vstup siete prezentujeme vstupné dáta ktoré sa preširujú k neurónom skrytej vrstvy, ktoré ich spracujú ako prechodovú funkciu váženej sumy vstupov a príslušajúcich váh. Je dôležité aby prechodová funkcia nebola lineárna, preto sa používa napríklad sigmoidálna funkcia, \tanh . Ďalšie vrstvy pracujú analogicky, teda spracovávajú vždy výstupy predchádzajúcej vrstvy. Výstupom neurónovej siete je výstup poslednej vrstvy.

Na trénovanie viacvrstvovej neurónovej siete sa môže použiť učenie so spätným šírením chyby [6], ktoré je založené na vypočítaní chyby na výstupe neurónovej siete a spätnej propagácii tejto chyby cez neurónovú sieť s úpravou váh.

Neurónová sieť s jednou skrytou vrstvou je univerzálny aproximátor funkcie [4]. Problémom ale je, že pri reálnych problémoch je pre dosiahnutie požadovanej presnosti potrebné použiť veľmi veľkú skrytú vrstvu. Veľkosť skrytej vrstvy ale prináša nárast počtu váh siete, pre ktoré treba trénovaním nájsť optimálne hodnoty, čo zvyšuje výpočtovú náročnosť ako aj požiadavky na veľkosť trénovacej množiny. Pri veľkých neurónových sieťach hrozí riziko pretrénovania, kedy sa neurónová sieť naučí trénovacie vzory ako keby naspamäť a stratí schopnosť generalizácie. Ukazuje sa, že hlboké neurónové siete s väčším počtom skrytých vrstiev dokážu nájsť riešenie efektívnejšie s menším počtom váh.

Zväčšovanie počtu skrytých vrstiev prináša dva problémy, prvým je možnosť pretrénovania a druhým je „vytrácajúci sa gradient“, ktorý pri použití spätného šírenia chyby spôsobí, že sa hodnota gradientu znižuje prechodom cez vrstvy a potom zmena váh v nižších vrstvách je minimálna.

Dobрым spôsobom ako týmto problémom čeliť je predtrénovanie vrstiev siete s použitím autoenkóderov a RBM (restricted Boltzman Machine) sietí [2]. Pri trénovaní neurónovej siete vzniká na skrytej vrstve vnútorná reprezentácia vstupu. Tento fakt sa dá použiť pri trénovaní autoenkóderov – sietí, ktoré majú rovnaký počet neurónov na

vstupnej aj výstupnej vrstve. Spravidla skrytá vrstva obsahuje menší počet neurónov. Pri tréningu sa na vstupe prezentujú vzory z tréningovej množiny a na výstupe požadujeme ten istý vzor. Nútime teda sieť, aby si na skrytej vrstve vytvorila reprezentáciu vstupu, ktorá ale vzhľadom na menší počet neurónov je v redukovanej dimenzii a teda sieť musí vyberať len podstatné črty. Týmto spôsobom sa dajú postupne natrénovať hlbšie siete, kedy pridávame ďalšie autoenkóдеры a spájame ich vo vrstvách za sebou. Keďže autoenkóдеры vytvárajú len vnútornú reprezentáciu dát, poslednou vrstvou pri sieťach typu vrstvených autoenkóderov (Stacked Autoencoder) je normálna výstupná vrstva doprednej siete a celá hlboká sieť sa dotrénuje prístupom spätného šírenia chyby.

Boltzmanove stroje sú neurónové siete schopné učenia sa bez učiteľa, čiže dokážu nájsť vzory v množstve vstupných dát bez toho, aby potrebovali vedieť, aký má byť výstup. Vytvoria vlastne lepšiu reprezentáciu dát na vyššej úrovni abstrakcie. Problémom Boltzmanových strojov však je, že aj keď majú veľmi dobre prepracovanú teóriu, v praxi nie sú veľmi dobre použiteľné, pretože ich učenie je príliš pomalé. Existuje však viacero spôsobov úpravy architektúry Boltzmanovho stroja a jeho procesu učenia, ktoré výrazne zefektívnia jeho učenie a vďaka tomu sú veľmi dobre použiteľné v praxi. Jednou z takýchto úprav Boltzmanovho stroja je Restricted Boltzmann Machine (RBM), čo znamená Boltzmannov stroj s obmedzením. Toto obmedzenie spočíva v tom, že RBM nemá žiadne prepojenia medzi dvomi skrytými neurónmi a dvomi viditeľnými neurónmi. Jeho architektúrou je teda bipartitný graf, kde sú skrytá a viditeľná vrstva neurónov vzájomne kompletne prepojené, no v rámci samotnej vrstvy nie sú žiadne prepojenia. Tak ako pri autoenkóderoch, môžeme aj RBM poskladať na seba a vznikne nám tzv. Deep Belief Network.

3 Príklady použitia hlbokých sietí

Jednou z oblastí v ktorých sme skúmali možnosti uplatnenia neurónových sietí sú úlohy strojového učenia nad sekvenčnými dátami. Jeden zo scenárov, na ktoré sme sa zamerali je hľadanie lepšej reprezentácie sekvencie dát zo zariadenia na snímanie pohľadu (eye-tracker). Použili sme sieť typu Restricted Boltzmann Machine (RBM), ktorej sme postupne ukazovali vizuálnu reprezentáciu používateľovho sedenia pred eye-trackerom v podobe teplotných máp, ktoré zachytávali priestorovú aj časovú informáciu o pohľade používateľa [1]. RBM sieť bola schopná nájsť také črty, ktoré poskytli vhodnú, abstraktnejšiu reprezentáciu jedného sedenia používateľa, ktorú sme úspešne použili pre ďalšie úlohy strojového učenia (segmentáciu aktivity počas sedenia).

Skúmali sme aj rekurentné neurónové siete: použitie LSTM siete [3] pre účely predikcie konverzie čitateľov článkov na webe s integrovaným platobným systémom na základe štandardných záznamov o prístupoch používateľov k obsahu. Skúmali sme vplyv architektúr, techniky odstavenia neurónov, miešania dát, vstupných údajov a aktivovania brány resetu na výkonnosť nášho modelu, ktorý sme porovnávali s náhodným lesom postaveným nad ľuďmi zvolenými črtami. Ukázalo sa, že naša architektúra so zapojením LSTM siete predikuje lepšie výsledky ako náhodný les (podľa F-metricky 17 % vs. 46 %), pričom nedosahuje rovnakú presnosť ako náhodný les, ale poráža ho v úplnosti.

V rámci výskumu extrakcie diskriminačných kľúčových slov [8] sme skúmali aj možnosť vytvorenia architektúry neurónovej siete, ktorá by bola ideálna pre túto úlohu. Väčšina architektúr neurónových sietí na extrakciu kľúčových slov bola doposiaľ štandardne navrhovaná tak, aby sme sieť natrénovali s učiteľom [9]. Pri takomto učení s učiteľom máme k dispozícii množinu textových dokumentov, pre ktoré sú známe ich kľúčové slová. My sme sa však zamerali na pokročilejšie techniky návrhu architektúr, ktoré by nevyžadovali dokumenty so známymi kľúčovými slovami. Namiesto toho sa zameriavame na úlohu kategorizácie dokumentov, v rámci ktorej nás zaujímajú diskriminačné kľúčové slová, t.j. také, ktoré majú dobrú rozlišovaciu schopnosť medzi danými kategóriami. Inšpirovaní Inception modulom [7] sme sa zamerali na návrh univerzálneho modulu pre extrakciu diskriminačných kľúčových slov. Základnou myšlienkou architektúry našej neurónovej siete je modelovanie kľúčových slov na medzivrstve, nie na výstupnej vrstve. Funkciou (v našom prípade dvoch) výstupných vrstiev je poskytnúť spätnú väzbu pre medzivrstvu kľúčových slov. Jedna výstupná vrstva zabezpečuje, aby črty medzivrstvy kľúčových slov reprezentovali črty skutočných slov, ktoré sa nachádzajú v príslušnom dokumente. Druhá výstupná vrstva zabezpečuje, aby črty medzivrstvy kľúčových slov boli diskriminačné, čo má tiež pozitívny vedľajší efekt, keďže na tejto výstupnej vrstve vidíme pravdepodobnosti zaradenia dokumentu do jednotlivých kategórií. Navrhnutú architektúru sa nám podarilo úspešne skombinovať s viacerými štandardnými architektúrami neurónových sietí ako sú konvolučné [5] a LSTM siete.

4 Záver

V našom príspevku sme v krátkosti predstavili niektoré modely hlbokých neurónových sietí, ako aj ich využitie pri rôznych úlohách strojového učenia. Súčasným trendom je kombinovanie rôznych typov sietí do vrstiev naozaj hlbokých a komplexných architektúr neurónových sietí. Existuje množstvo rámcov (angl. frameworkov) ako napr. TensorFlow¹, Torch² alebo Theano³, ktoré umožňujú jednoduchú tvorbu takto vrstvených sietí a ich tréning čipoch grafických kariet. Myslíme si, že takýmto kombinovaním a striedaním rôznych typov sietí vo vrstvách hlbokých sietí, bude možné dosahovať ešte lepšie výsledky pri úlohách napodobňujúcich ľudskú inteligenciu.

Podakovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektov: Prispôbovanie prístupu k informačným a vedomostným artefaktom založené na interakciách a kolaborácii v prostredí webu (VG 1/0646/15), Inteligentná analýza veľkých údajových korpusov sémanticky-orientovanými a bio-inšpirovanými metódami v paralelnom prostredí (VG 1/0752/14), Informačné správanie sa človeka v digitálnom priestore (APVV-15-0508) a projektu v rámci OP Výskum a vývoj pre projekt: Medzinárodné

¹ TensorFlow – Open Source library for Machine Intelligence , <https://www.tensorflow.org/>

² Torch – scientific computing on GPUs, <http://torch.ch/>

³ Theano – python library for deep learning, <http://deeplearning.net/software/theano/>

centrum excelentnosti pre výskum inteligentných a bezpečných informačnokomunikačných technológií a systémov, ITMS 26240120039, spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja.

Literatúra

1. Barla, M., Šimek, M., Bieliková, M.: Comparing Eye-tracking Data Using Machine Learning. In: Journal of Eye Movement Research, Vol. 8, No. 4 (Special Issue ECEM 2015). Abstract. p. 192.
2. Hinton, G. E., Salakhutdinov, R. R.: Reducing the Dimensionality of Data with Neural Networks. In: Science, Vol. 313 No. 5786, pp. 504–507, 2006.
3. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. In: Neural Computation Vol. 9 No. 8, pp. 1735–1780, MIT Press, 1997.
4. Hornik, K.: Approximation capabilities of multilayer feedforward networks. In: Neural Networks, Vol. 4, No. 2, pp. 251–257, Elsevier, 1991.
5. LeCun, Y., et al.: Gradient-based learning applied to document recognition. In: Proceedings of the IEEE, Vol. 86, pp. 2278–2324, 1998.
6. Rumelhart, D.E., Hinton, G.E., Williams, R.J. Learning representations by back-propagating errors. In: Nature, Vol. 323, pp. 533 – 536, 1986.
7. Szegedy, C. et al.: Going deeper with convolutions. In: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR 2015, pp. 1–9, IEEE, 2015.
8. Šajgalík, M., Barla, M., Bieliková, M.: Exploring Multidimensional Continuous Feature Space to Extract Relevant Words. In: Proceedings of Statistical language and speech processing: second international conference, SLSP 2014, pp. 159–170, Springer, 2014.
9. Taeho, J., Lee, M., Gatton, T.M.: Keyword extraction from documents using a neural network model. In: Proceedings of the 2006 International Conference on Hybrid Information Technology, Vol. 2, pp. 194–197, IEEE, 2006.

Annotation:

Solving (Non-)standard Machine Learning Tasks by Deep Learning

Big international companies such as Google, Microsoft or Facebook invest fair amounts of resources into R&D in the field of deep learning. Recent defeat of Go game champion by a deep neural network shows potential of this approach. Applications based on deep neural networks beat models based on human crafted and fine-tuned features in many different domains. In this invited talk, we are going to explain basic principles of deep learning and show its application on different problems being solved at IISE FIIT SUT.

R vs. Python – which one fits you best?

Jakub Ševcech, Peter Laurinec, Ondrej Kaššák

Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava, Slovenská republika

{jakub.sevcech,peter.laurinec,ondrej.kassak}@stuba.sk

Abstract. Both Python and R are popular programming languages for data analysis. While R's functionality is developed with statisticians in mind, Python is often praised for its easy-to-understand syntax. In this paper, we will highlight some of the differences between R and Python, and how they both have a place in the data science world.

Paper type: Invited talk

Keywords: R, Python, Data Analysis, Data Science

1 Introduction

R or Python, which to choose? Which is better? These questions worry researchers for a long while without clear or obvious result. As data science has many faces, there exist fields where it is more suitable to use R and other use cases, where Python is the language to choose. The aim of our paper is not to create another paper which ends with a diplomatic tie. We rather focus on description of practical experiences and giving the advices how to choose language for concrete researcher and research task.

2 Language characteristics

If you take a look at communities of both languages, their activity on github and discussion forums, you find that both of them are numerous and active. The R is popular mainly in academic environment and it is slowly moving into the enterprise sphere. Python community is, on the other hand, formed mostly by engineers, programmers and hackers, who moved into the data analysis field. Nowadays it appears that R is for a head in advance in fields like time series analysis, econometrics, robust statistical methods, bayesian statistics and machine learning, but in our opinion it is mostly by a historical reasons. Few years back (and also today in some extent), R had more packages for statistical analysis, data processing and machine learning than Python.

However, by actively borrowing the good parts from various mathematical languages, Python library offer is growing rapidly. It introduced the support for data frame

manipulation from R into Python in the Pandas library, numerical analysis and matrix operation best known from languages as Matlab or Octave in library Numpy, many machine learning algorithms in scikit-learn and many more. These libraries are actively developed and rapidly growing in an organized way as they can build on years of experience from other, more mature languages.

An argument in favor of R is that even if the Python is easy to learn, you need to get your hand dirty and write more interlinking code between various parts from these libraries. In R, you follow the typical workflow optimized for typical steps and the analysis is straightforward, generally just in a few lines of code. Hence, the R language is very easy to use. However, this can often come back to bite you, if you have to do something atypical or create something never done before. On the other hand Python offers full-valued alternative for the data science and it brings advantages of programming language with uniform syntax, high code readability and testability.

3 Which language fits my research the best

The process of selection between R and Python should subject mainly to the task, it will be used for. The closer the task is to the pure mathematics and research, the more suitable it is to use the R language with its data frames, matrix data processing, machine learning, statistical testing and visualization packages. On the other side, in the case of more real world problems with need to work with messy data from multiple sources and when applying research result into production, Python is more suitable.

In case you are choosing the suitable language for your research, you should consider what is your aim. If you are using big amounts of data, you need to easily evaluate various algorithms and the result of your research will be an analysis or research paper, R is a logical choice. You can achieve this result without the need of extensive coding, testing and creation of production application. You simply use existing libraries for data manipulation, analysis and visualization and build the script in one go.

On the other side, if you need to collect the data at first, crawl various web sources with different structure, process the data and then create a production service from it, you should choose rather Python. It is also more suitable for projects where bigger codebase will be built, maintained and read multiple times, potentially by different people. The main aim of Python language authors was the code beauty and readability and as the codebase grows, it tends to be much more maintainable when it is written in Python than in R. The R language is more suitable for individual analyses or algorithms and not so much for complex systems.

4 Typical data science workflow

The typical workflow when using the R language closely follow the typical progress of any data analysis. You load the data, preprocess it, visualize, learn the model, visualize it, its results and statistically evaluate its accuracy. Of course, this is a very limited characteristic and one can find other modes of application of R spiraling out of this typical one.

In the case of Python, you can not form such stable workflow as it was designed as general purpose language and its uses are really broad. It can be used in the data collection phase, in data transformation in novel manners, it can be used in the typical data analysis workflow (even though, the typical roads are not so well-worn and you need more interlinking code to join various libraries), but it can be also used to build complex systems and production applications employing data analysis results. In general, Python is very flexible and you, as a data scientist, can benefit from this when doing something novel that has never been done before.

5 Conclusions

The more tools you have at your disposal the more effectively you can do the job at hand. So the best choice is to learn both and choose the most suitable tool for the problem you are currently dealing with.

Another good advice is to master yourself in some language. It can be anything - R, Python or even Java. Only if you know some language well, you are able to use it effectively and benefit from its features and strengths.

Both are getting better at doing what the other does well though. We already noted that Python has Pandas to mimic R functionality. R has a web application framework called Shiny. There are libraries to use R with Python, and vice versa. We'd just recommend using both. When you need something that is general purpose Python is better. When you just need to do data analysis or answer a question, R is better.

Acknowledgment: This work was partially supported by the Research and Development Operational Programme for the project International Centre of Excellence for Research of Intelligent and Secure Information-Communication Technologies and Systems, ITMS 26240120039, the Scientific Grant Agency of The Slovak Republic, grant No. VG 1/0752/14 and VG 1/0646/15, the Slovak Research and Development Agency under the contract No. APVV-15-0508 and the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 009STU-4/2014.

**Analýza dát,
dolovanie v dátach
a strojové učenie**

Course Similarity Analysis

Hana Bydžovská

CSU and KD Lab Faculty of Informatics
Masaryk University, Brno

`bydzovska@fi.muni.cz`

Abstract. Courses offered to students at universities have different characteristics. In this paper, we analyse course similarities to improve the students' performance prediction. We utilize the item-to-item collaborative filtering approach that computes course similarities based on students' grades. We also use content based techniques to compute course similarities based on the information from the course catalogue, e.g. the course content or prerequisites. Using the computed similarities and utilizing different clustering algorithms, we are able to reveal interesting course groups that can be used to improve the student performance prediction. Finally, we are able to predict the students' final grades of the investigated course by examining grades of only three related courses.

Contribution type: Application paper

Keywords: course similarity, student performance prediction, university information system

1 Introduction

The problem of the student grade prediction in a particular course has recently been addressed using data mining techniques. Researchers usually examine study-related records, e.g. the age, the gender, and the field of study [5] because of their easy availability in university information systems. Moreover, they attempt to identify additional characteristics that can lead to better understanding of students' behaviour, e.g. their habits [3] or parents' education [6]. The most typical way how to obtain such data is to conduct questionnaires. We cannot rely on data obtained by questionnaires since they tend to have a lower response rate. Therefore, only the data originated from the Information System of Masaryk University are employed for our experiments. Our approach is based on recommender system techniques [2] applied to the educational context. We mapped the users-item-rating problem to the student-course-grade problem and predict the final grades based on previous achievements of similar students. We also succeeded in identifying course dependencies. Finally, we were able to predict the final grades of the investigated course by examining grades of only 3 other courses.

2 Course Similarity

2.1 Students' Grades

The collaborative filtering item-to-item approach from the recommender system theory was utilized. We mapped the users-item-rating problem to the student-course-grade problem [1]. The first step was to construct a matrix G where rows represented students and columns represented courses. Grades formed the matrix. If a student did not attend a particular course, the corresponding cell remained empty. The adjusted cosine measure is then calculated from the previously defined matrix G for each pair of courses.

2.2 Course Characteristics

Students search for useful information about courses in the Course Catalogue that help them to decide whether or not they should enrol the course. We selected different course characteristics and attempted to identify dependencies among courses [1]. Similarity of courses was defined by the weighted sum of the similarities of the selected course characteristics: prerequisites, literature, course content represented by the text about the study subject and outline, teachers, and course supervisor.

3 Course Clustering

Subsequently, we could construct a similarity matrix for each previously mentioned approach where rows and columns represented courses. The value defined similarity among courses formed the matrix. For both matrices, we utilized three different clustering algorithms to create course clusters: k-mean, spectral clustering and average link clustering [4]. The resulted clusters defined the groups of similar courses.

For each clustering settings, we also computed Davies–Bouldin index and Dunn index to assign the best score to the algorithms with their settings that produces clusters with high similarity within a cluster and low similarity between clusters.

To be able to analyse created course groups, we designed application (see Figure 1) that allows us to visualize course groups. The application can also help university management to revise course characteristics, their difficulties, or location in course templates that define students study plans.

4 Student Performance Prediction

For grade prediction, we utilize collaborative filtering user-to-user approach. We used matrix G defined in Section 2.1 and students' similarity was calculated by Pearson's Correlation Coefficient. Finally, when we predicted the students' grades of a certain course, we reduced the computations to the grades obtained from courses belonging to the same cluster as the investigated course.

Fig. 2. Application for Faculty Management.

In this section, we focused on clusters obtained by hierarchical clustering algorithm [4]. In comparison with the method using all grades, both approaches (similarity using grades: SC_1 , similarity using course characteristics: SC_2) had positive effects on the number of calculations (see Table 1). 123 courses from all 138 belonged to some of the created clusters and the final grades could be predicted based on the grades of only 3 other courses on average. 70 of our investigated courses belonged to different clusters using SC_1 and SC_2 . A slightly better MAE was obtained by the method utilizing the course characteristics for these courses. Therefore, when a grade is predicted, the corresponding course is searched in SC_2 , then SC_1 .

Table 1. Comparison of SC_1 and SC_2

| Method | MAE | Sensitivity | Number of clusters | Average cluster size | Shared Courses |
|------------|-------|-------------|--------------------|----------------------|----------------|
| All grades | 0.687 | 0.402 | 1 | 499 | 10 |
| SC_1 | 0.681 | 0.390 | 37 | 3 | 3 |
| SC_2 | 0.640 | 0.386 | 36 | 3 | 2 |

5 Conclusion

In this paper, we focused on the problem of predicting final grades of students. Our approach utilized recommender system techniques and predicted grades based on the similarity of students' achievements. Each university information system stores the data about students' grades which were needed for the prediction. We also succeeded in identifying course dependencies. Finally, we were able to predict the final grades of the investigated course by examining grades of only 3 other courses.

Once we have a reliable performance prediction, it can be used in many contexts: for identifying weak students, for guiding the adaptive behaviour in intelligent tutoring systems, or for providing a feedback to students. We are interested in designing a course enrolment recommender system that will help students with selecting courses to enrol in.

References

1. Bydžovská H.: A Comparative Analysis of Techniques for Predicting Student Performance. In Proceedings of the 9th International Conference on Educational Data Mining (EDM'16), pp. 306-311, (2016).
2. Manouselis, N. and Drachsler, H. and Vuorikari, R. and Hummel, H. and Koper, R.: Recommender Systems in Technology Enhanced Learning, Recommender systems Handbook Springer Verlag 2011, pp 387-415, (2011).
3. Marquez-Vera, C. Romero, C. and Ventura, S.: Predicting school failure using data mining. In Proceedings of the 4th International Conference on Educational Data Mining (EDM'11), pp. 271-276, (2011).
4. Murtagh, F. and Contreras, P.: Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86-97, (2012).
5. Nghe, T. N., Janecek, P. and Haddawy, P.: A comparative analysis of techniques for predicting academic performance. 37th ASEE/IEEE Frontiers in Education Conference, Milwaukee, WI 2007, (2007).
6. Vandamme, J.P., N. Meskens and Superby, J. F.: Predicting academic performance by data mining methods. *Educ. Econ.*, 15, 405-419, (2007).

Detekcia nebezpečných aktivít v záznamoch udalostí mobilných zariadení

Štefan Dlugolinský, Giang Nguyen a Ladislav Hluchý

Ústav informatiky, Slovenská akadémia vied
Dúbravská cesta 9, 845 07 Bratislava

{stefan.dlugolinsky, giang.ui, ladislav.hluchy}@savba.sk

Abstrakt. V článku prezentujeme experiment, ktorý sme vykonali v rámci projektu venujúceho sa bezpečnosti mobilných zariadení. V experimente sme sa pokúsili aplikovať metódu modelovania jazyka pomocou n-gramov na doménu bezpečnosti mobilných zariadení. Cieľom bolo zistiť, či je možné použiť n-gram modely vytvorené zo záznamov udalostí mobilných zariadení na odhaľovanie nebezpečných udalostí a reťazcov udalostí.

Typ príspevku: Výskumný príspevok

Kľúčové slová: bezpečnosť, n-gramy, modelovanie jazyka

1 Úvod

Predmetom našich experimentov bolo preskúmanie využitia metód pravdepodobnostného modelovania jazyka (v angl. literatúre ako Probabilistic Language Modelling) v doméne bezpečnosti mobilných zariadení. V experimentoch sme namiesto postupností slov jazyka modelovali postupnosti udalostí zachytených v záznamoch mobilných zariadení (telefóny a tablety so systémom Android). Tak ako v prirodzenom jazyku, tak aj v záznamoch udalostí sme predpokladali určitú závislosť nasledujúcej udalosti od predchádzajúcich udalostí. V prirodzenom jazyku táto vlastnosť predstavuje sémantiku, kde určitá postupnosť slov dáva nejaký význam. V záznamoch udalostí sme predpokladali, že to je časť nejakého procesu pozostávajúceho z určitých akcií, ktoré sú zaznamenané ako sled udalostí. Našou snahou bolo vytvoriť modely nebezpečných reťazcov udalostí zo záznamov udalostí a využiť vytvorené modely na detekciu podozrivej aktivity v mobilných zariadeniach. V experimente sme aplikovali modely na vzorky udalostí s podozrivou aktivitou ako aj bez nej. Výsledky sme vyhodnotili metrikami perplexity a logaritmickej pravdepodobnosti (pravdepodobnosť, s akou sa testovaná vzorka podobá na vzorky z trénovacej množiny modelu).

2 Prehľad súčasného stavu

Podľa našich zistení sme nenašli literatúru, ktorá by sa zaoberala využitím metód pravdepodobnostného modelovania jazyka na odhaľovanie podozrivých reťazcov zo záznamov udalostí mobilných zariadení. Podobné prístupy však možno nájsť v práci [1], kde autori testovali presnosť detekcie podozrivých častí kódu pomocou n-gram modelov. Predbežné výsledky ukazovali 98% presnosť detekcie pri 3-násobnej krížovej validácii na datasete pozostávajúcom zo 65 podozrivých programov získaných z emailovej komunikácie. V ďalšej príbuznej práci [3] sa autori zaoberali využitím n-gram modelov na rozpoznávanie neznámych malware v súvislosti s metódou signature-based detection. Ich výsledky ukázali, že n-gram modely dokážu detekovať aj neznáme vzorky kódu.

3 Modelovanie prirodzeného jazyka

Pravdepodobnostné modelovanie jazyka je známe z oblasti spracovania prirodzeného jazyka. Často sa využíva na riešenie rôznych úloh ako napríklad: a) detekcia jazyka, b) automatická korekcia chýb v texte, c) predikcia pri písaní textu, d) strojový preklad a e) rozpoznávanie písaného textu. Princíp spočíva vo vytvorení pravdepodobnostného modelu, ktorý reprezentuje rozdelenie pravdepodobnosti všetkých možných reťazcov slov daného jazyka. Na základe rozdelenia pravdepodobnosti je možné určiť mieru príslušnosti vstupného textu k namodelovanému jazyku, pričom sa berú do úvahy závislosti po sebe idúcich slov v reťazcoch. Tradične sa na modelovanie jazyka používa metóda n-gramov, teda n-tíc slov, ktoré možno pravidlami modelovaného jazyka vytvoriť. N-gramy sa na vytvorenie modelu získajú z trénovacieho textu/trénovacej množiny. Pod pojmom model jazyka rozumieme rozdelenie pravdepodobnosti nad reťazcami trénovacej množiny, pričom model vyjadruje pravdepodobnosť s akou vstupný reťazec predstavuje vetu modelovaného jazyka. Napr., pravdepodobnosť reťazca r dĺžky d pozostávajúceho zo slov $r_1 r_2 \dots r_d = r_1^d = r$ môžeme vyjadriť pomocou vzťahu (1).

$$P(r) = \prod_{i=1}^d P(r_i | r_1 \dots r_{i-1}) \quad (1)$$

Výhodnejšie je pravdepodobnosť aproximovať tak, že pravdepodobnosť nasledujúceho slova závisí od slova alebo reťazca slov pred ním. Podľa toho stupňujeme aj modely. Napr. bi-gram model aproximuje pravdepodobnosť nasledujúceho slova na základe predchádzajúceho slova; vzťah (2). Tri-gram model zasa na základe dvojice predchádzajúcich slov; vzťah (3). Analogicky takto aproximujeme pravdepodobnosť aj pre modely vyššieho stupňa.

$$P(r) \approx \prod_{i=1}^d P(r_i | r_{i-1}) \quad (2)$$

$$P(r) \approx \prod_{i=1}^d P(r_i | r_{i-2} r_{i-1}) \quad (3)$$

Dôležitým krokom pri vytváraní n-gram modelu je odhad pravdepodobnosti pre známe n-gramy trénovacieho datasetu. Najjednoduchším spôsobom je odhad pomocou frekvencie n-gramov v datasete (v angl. literatúre Maximum-likelihood estimate). Problémom prístupu n-gramov však je, že najlepšie fungujú vtedy, ak je testovacia množina podobná tej trénovacej. V praxi to však býva niekedy problém. Keďže trénovací dataset

nemôže pokryť všetky možné n-gramy, neznáme n-gramy dostanú pri takomto odhade nulovú pravdepodobnosť – problém riedkeho datasetu. Preto sa môže stať, že ak sa nám neznámy n-gram objaví v testovacom datasete, tak mu model priradí nulovú pravdepodobnosť a vyhodnotí že tento n-gram nepatrí do modelovaného jazyka. Tento problém sa rieši vyhladzovaním odhadu pravdepodobnosti s cieľom priradiť neznámym n-gramom určitú malú pravdepodobnosť. Na vyhladzovanie pravdepodobnosti poznáme niekoľko metód [2]:

a) Add-one smoothing, b) linear interpolation, c) Good-Turing smoothing, d) Jelinek-Mercer smoothing, e) Katz (backoff), f) Witten-Bell smoothing, g) Absolute discounting a h) Kneser-Ney smoothing. Pre potreby našich experimentov sme zvolili algoritmus Kneser-Ney, ktorý sa najlepšie osvedčil v doméne modelovania prirodzeného jazyka. Použili sme pôvodnú verziu s interpoláciou.

4 Dáta

Dáta vo forme záznamov udalostí pochádzali z 18 mobilných zariadení používaných testovacími subjektami a boli zbierané počas obdobia troch mesiacov. Zozbierané záznamy vo forme vektorov s atribútmi sme spracovávali na distribuovanom úložisku tvorenom Hadoop klastrom s nástrojmi Apache Pig a Apache Hive. Zo zariadení sa zbierali dáta o a) uskutočnených hovoroch (CALLS), b) SMS komunikácii (SMS), c) systémových volaniach (INTENT_RECEIVED), d) informáciách o spustených procesoch (PROCESSES), e) sieťovej komunikácii (CONNECTIONS) a f) histórii webového prehliadača (BROWSER_HISTORY).

4.1 Analýza

Zo zozbieraných záznamov sme náhodným výberom vybrali približne 9 mil., ktoré sme podrobili frekvenčnej analýze. Charakteristika vybranej vzorky dát je v Tab. 1.

Tab 1. Frekvenčná analýza vybranej vzorky dát.

| typ udalosti | zariadenia | záznamy | unikátne záznamy |
|-----------------|------------|-----------|------------------|
| SMS | 10 | 3 231 | 462 14.30% |
| CALLS | 11 | 3 187 | 651 20.43% |
| INTENT_RECEIVED | 18 | 729 332 | 572 353 78.48% |
| PROCESSES | 18 | 4 372 613 | 3 992 536 91.31% |
| CONNECTIONS | 18 | 1 951 405 | 1 797 525 92.11% |
| BROWSER_HISTORY | 15 | 1 919 961 | 1 098 979 57.24% |
| celkovo | 18 | 8 979 729 | 7 462 506 83.10% |

Frekvenčná analýza pozostávala z vytvorenia štatistiky o hodnotách jednotlivých atribútov. Cieľom bolo identifikovať také atribúty, ktoré špecifikujú určitý typ udalosti, aby sme pomocou nich mohli udalosti transformovať na všeobecnejšie slová. Napr. udalostí typu CONNECTION bolo až 92% jedinečných, teda sa skoro vôbec neopakovali. Dôvodom bola široká škála hodnôt atribútov ako aj počet samotných atribútov. Potrebovali sme preto odstrániť niektoré atribúty, prípadne kategorizovať ich hodnoty.

Z udalostí, ktoré obsahovali atribúty: *time*, *gps*, *acc*, *from_addr*, *from_port*, *to_addr*, *to_port*, *state*, *uid*, *application*, *protocol* a *imei* sme ponechali len atribúty *application*, *to_port*, *protocol* a *state*. Tie sme potom transformovali na slová reprezentujúce udalosti v tvare: **connection://{application}/{to_port} /{protocol}/{state}**. Výber atribútov bol urobený na základe odporúčaní experta na zozbierané dáta.

4.2 Predspracovanie

Ako bolo spomenuté v predchádzajúcom príklade, udalosti sme filtrovali a transformovali z vektorov na slová (event), pričom každé slovo malo aj svoju časovú značku (time). Transformácia prebehla podľa nasledovnej schémy:

```
time          event
YYYY-MM-dd HH:mm:ss.SSS  type://attr_1/attr_2/.../attr_n
```

kde *type* bol typ udalosti (napr. BROWSER_HISTORY pre udalosť z webového prehliadača) a *attr₁*, *attr₂* až *attr_n* hodnoty vybraných atribútov (napr. protokol). Udalosti transformované na slová sme ďalej spájali do sekvencií, ktoré boli ekvivalentné vetám v prirodzenom jazyku a udalosti v sekvenciách zasa slovám vo vetách. Udalosti sme spájali tak, aby časový rozdiel medzi dvoma po sebe idúcimi udalosťami v sekvencii nebol väčší ako 10 sekúnd. Túto hodnotu sme zvolili po predchádzajúcej diskusii s expertom na mobilné zariadenia. Spájanie udalostí do sekvencií nám umožňovalo neskôr generovať n-gramy a z nich potom pravdepodobnostný model reťazcov udalostí, podobne ako by to bolo pri texte.

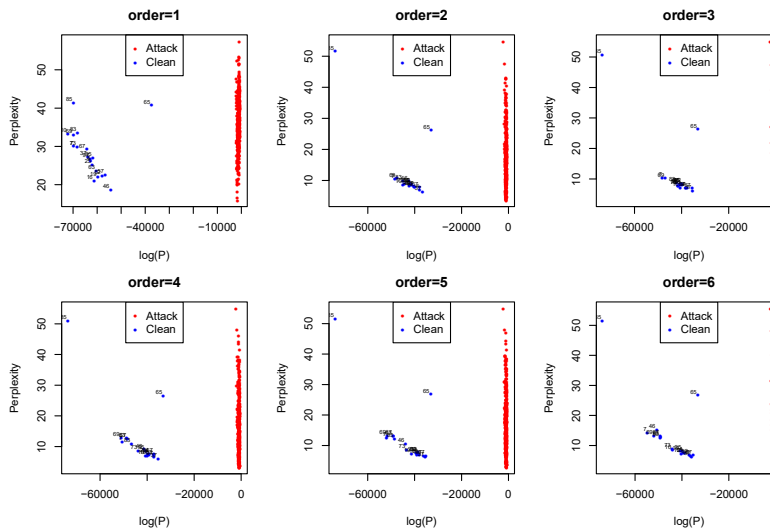
4.3 Vzorky útokov

Pomocou experta boli vykonané simulácie dvoch typov útokov: a) Útok1 - získanie citlivých údajov (Cookies, uložené heslá, auto-fill dáta) a b) Útok2 - získanie vzdialeného prístupu na shell napadnutého zariadenia. K dispozícii sme mali 97 vzoriek útokov typu 1 (59) a 2 (38), ktoré sme zo záznamov získali podľa časových intervalov začiatku a konca simulácie útokov. Vzorky mali v priemere okolo 400 slov (udalostí).

5 Realizácia experimentu

Na odporúčanie experta, ktorý vykonával simulácie útokov, sme si zvolili útok typu 2 a pre každú vzorku útoku tohto typu sme vygenerovali n-gram modely stupňa 1 až 6 s vyhladzovaním pravdepodobnosti algoritmom Kneser-Ney. Pre každý z 38 útokov typu 2 sme tak mali 6 modelov, ktoré sme následne vyhodnocovali nad dvoma typmi datasetov: **CLEAN** Datasets udalostí bez podozrivej aktivity – Vytvorili sme jeden dataset zo záznamov zariadenia A v období štyroch dní. Celkovo obsahoval 47 132 udalostí v 901 sekvenciách. Tento dataset sme používali ako testovací dataset s bezpečnými udalosťami. **ATTACK** Datasets udalostí s útokom typu 2 – Pre každú z 38 vzoriek útoku typu 2 sme vytvorili jeden dataset. Každý dataset bol tvorený sekvenciami udalostí, kde dve udalosti nasledujúce v sekvencii po sebe neboli od seba v čase vzdialené viac ako 10 sekúnd. Z týchto 38 datasetov bolo 19 datasetov zo záznamov zariadenia A. Ostatné boli zo zariadenia B (15 útokov) a z C (4 útoky). Pre potreby experimentu sme si vybrali 19 datasetov zariadenia A, z ktorých sme vygenerovali rovnaký počet n-gram modelov

pre každý jeden stupeň (1 až 6). Tieto modely ďalej spomínáme ako *attack* modely. Zvyšné datasety a tiež aj týchto 19 vybraných sme použili ako testovacie datasety s podozrivými udalosťami. Keďže sme testovacie datasety CLEAN a ATTACK vytvorili zo zariadenia, na ktorom boli simulované útoky, overili sme si, že sa nám podozrivé udalosti nedostali do CLEAN datasetu. Evaluáciu *attack* modelov sme vykonali v dvoch krokoch. V prvom kroku sme *attack* modely evaluovali nad všetkými ATTACK datasetmi okrem tých, z ktorých boli modely vytvorené (nezávislosť od tréningového datasetu). Vykonali sme tak spolu 342 evaluácií pre každý stupeň n-gramu (1 až 6) a sledovali sme metriky perplexity a logaritmickej pravdepodobnosti príslušnosti reťazcov datasetu k namodelovaným útokom. V druhom kroku sme evaluovali *attack* modely nad CLEAN datasetom. Rovnako ako v prvom kroku, tak aj v druhom kroku sme sledovali metriky perplexity a logaritmickej pravdepodobnosti. Namerané hodnoty získané v oboch krokoch sme zobrazili v grafe na Obr. 1. Ako je vidieť z obrázka, hodnoty logaritmickej pravdepodobnosti získané evaluáciou *attack* modelov nad ATTACK datasetmi (krok 1 - červená farba) dosahovali vyššie hodnoty pravdepodobnosti ako pri evaluácii nad CLEAN datasetom (krok 2 - modrá farba; číslami sú podľa vzorky útoku označené *attack* modely). V prípade sledovanej metriky perplexity to už také jednoznačné nebolo. Výsledkom však bolo, že pomocou logaritmickej pravdepodobnosti by bolo možné odlíšiť podozrivé vzorky záznamov udalostí od bežných.



Obr. 1 Výsledky evaluácie attack modelov nad CLEAN a ATTACK datasetmi

6 Záver

Prvotné výsledky vykonaných experimentov ukazujú, že by bolo možné využiť metódy modelovania prirodzeného jazyka aj v oblasti bezpečnosti mobilných zariadení. V experimente sme evaluáciou modelov škodlivých udalostí získali výrazne vyššie hodnoty

logaritmickej pravdepodobnosti pre sekvencie udalostí so škodlivou aktivitou ako bez nej. Toto pozorovanie by sa mohlo využiť napr. na natrénovanie binárneho klasifikátora sekvencií udalostí, ktorý by bol s určitou mierou pravdepodobnosti schopný dekekovat' nebezpečné aktivity v mobilnom zariadení.

Pod'akovanie: Táto publikácia bola podporená projektami VEGA 2/0167/16 a EGI-Engage EU H2020-654142. Zároveň by sme sa chceli poďakovať všetkým kolegom, partnerom a doménovým expertom, ktorí s nami spolupracovali a diskutovali.

Literatúra

1. Abou-Assaleh, T., Cercone, N., Keselj, V., Sweidan, R.: N-gram-based detection of new malicious code. In: Computer Software and Applications Conference, 2004. COMPSAC 2004. Proceedings of the 28th Annual International. vol. 2, pp. 41–42 vol.2 (Sept 2004)
2. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. In: Proceedings of the 34th Annual Meeting on Association for Computational Linguistics. pp. 310–318. ACL '96, Association for Computational Linguistics, Stroudsburg, PA, USA (1996), <http://dx.doi.org/10.3115/981863.981904> ^[1]_{SEP}
3. Santos, I., Peña, Y.K., Devesa, J., Bringas, P.G.: N-grams-based file signatures for malware detection. In: ICEIS 2009 - Proceedings of the 11th International Conference on Enterprise Information Systems, Volume AIDSS, Milan, Italy, May 6-10, 2009. pp. 317–320 (2009)

Annotation:

Detection of malicious activity in the event logs of mobile devices

In this paper, we present experiment conducted within a mobile security project. We tried to apply methods of Natural Language Modelling in the domain of mobile device security. The point was to investigate, whether n-gram models created from event logs of mobile devices can be used to detect malicious events or sequences of such events.

Použitie transformačnej regresnej techniky pre dolovanie v údajoch

Peter Krammer, Ladislav Hluchý

Ústav Informatiky
Slovenská Akadémia Vied
Dúbravská cesta 9, 845 07 Bratislava, Slovenská republika

{peter.krammer, ladislav.hluchy}@savba.sk

Abstrakt. Metódy dolovania a strojového učenia je možné aplikovať v mnohých doménach. Avšak viaceré spomedzi oblastí generujú len obmedzený objem dát, resp. získanie väčšieho objemu dát je drahé, časovo resp. technicky náročné. Aj v týchto oblastiach však vzniká potreba modelovania a predikcie, pričom nižší objem dát spôsobuje problematické zovšeobecnenie vlastností, čo sa prejaví nižšou presnosťou modelu. Článok pojednáva o dátovej transformácii, ktorá klasickú regresnú úlohu transformuje na úlohu s výrazne vyšším počtom záznamov s cieľom zvýšenia presnosti modelovania. Článok prináša modifikáciu dátovej transformácie ako aj jej otestovanie na reálnych dátových množinách. Pri tom porovnáva a hodnotí dosiahnutú výkonnosť natrénovaných modelov.

Typ príspevku: Výskumný príspevok

Kľúčové slová: dátová transformácia, regresia, modelovanie, dolovanie

1 Úvod

Súčasným trendom v mnohých častiach informatiky je rozhodne oblasť veľkých dát. Táto oblasť poskytuje výzvu pri riešení problémov efektívneho narábania so zdrojmi, škálovateľnosti, ako aj použitia vhodnej distribuovanej architektúry a podobne. Pri dolovaní v údajoch sa však často stretávame s opačným prípadom, kedy nemôžeme hovoriť o veľkých dátach; pri dostupnosti len niekoľko tisíc záznamov či dokonca menej. Takéto prípady nastávajú v doménach, kde meranie a zbieranie dát je časovo, alebo technicky náročné, resp. ekonomicky nákladné. Problémom sa tak skôr stáva reprezentatívnosť dátovej množiny a schopnosť generalizovania vzťahov modelom, čo sa prejaví znížením miery presnosti modelu. Aj napriek týmto problémom však vzniká potreba modelovania a predikcie veličín aj z týchto oblastí. V súčasnosti sa za účelom spresňovania modelov zvyčajne používajú metódy združeného učenia [1]. Tieto metódy často zlučujú viaceré rozdielne typy modelov, ktoré tak vzájomne kompenzujú svoje slabé stránky. Výsledný združený model, ktorý je zložený z čiastkových modelov

tak obvykle dosahuje vyššiu mieru presnosti predikcií, pretože predpovede sú výsledkom hlasovania čiastkových modelov. Druhým často používaným princípom v združenom učení¹ je viacnásobné tréovanie jedného typu modelu, pričom váhy jednotlivých záznamov sa menia v závislosti od úspešnosti predpovedí. Tento spôsob používa aj známa metóda AdaBoost. Viaceré spomedzi metód združeného učenia (Boosting, Bagging) boli pôvodne určené pre úlohu klasifikácie do tried, avšak neskôr boli navrhnuté aj modifikácie pre úlohu regresie [2], [6]. Celkovo však tieto metódy vychádzajú z princípu zlúčenia viacerých modelov, čo vedie k zložitejšej štruktúre vytvoreného modelu. V našom príspevku však používame len jeden model, ktorý je tréovaný na transformovaných dátach. Ďalšími výhodami tohto prístupu sú možnosť použitia dodatočného združeného učenia (pre ďalšie spresnenie), ako aj možnosť voľby typu použitého modelu.

1.1 Základný princíp transformácie

Základnou ideou transformačnej techniky je transformovať pôvodnú regresnú úlohu na ekvivalentnú s vyšším počtom záznamov a atribútov tak, aby tieto dáta boli svojou štruktúrou vhodnejšie pre proces strojového učenia. Za týmto účelom je použitá dátová transformácia, ktorá z pôvodnej dátovej množiny postupne vyberá všetky možné dvojice záznamov, pričom jedna dvojica vytvára jeden záznam transformovanej dátovej množiny. Z pôvodných N záznamov v dátovej množine získame $N^2 - N$ záznamov v transformovanej množine. Počet vstupných atribútov sa zvýši dvojnásobne, keďže okrem príslušného atribútu bude zastúpená aj diferenciencia príslušného atribútu. Prezentovaná transformácia je vhodná v prípade úlohy regresie, výhradne pri spojitých numerických atribútoch, obzvlášť v prípade menších dátových množín.

Uvažujeme, že vstupné dáta, s homogénnou štruktúrou - v tvare tabuľky už boli predspracované a obsahujú vybrané relevantné vstupné atribúty. Po aplikovaní dátovej transformácie budú tieto údaje transformované do štruktúry obsahujúcej okrem pôvodných hodnôt aj ich diferenciencie. Cieľovou veličinou sa stane diferenciencia z pôvodných cieľových veličín. Pri predikcii je nutné opätovne realizovať dátovú transformáciu na predikovaný záznam, ktorý spárujeme so záznamami z tréovacej množiny. Získame tak väčší počet odhadov cieľovej veličiny, z ktorých následne určíme finálnu hodnotu cieľovej predikovanej veličiny. Podrobnejšie je dátová transformácia, jej základné aspekty, stratégie určenia cieľovej hodnoty ako aj proces predikcie popísané v publikácii [5].

Princíp ktorý umožňuje, aby táto technika dosahovala zlepšenie má niekoľko aspektov. V prvom rade, použitie rozdielov (diferencií) do určitej miery vyjadruje mieru vzdialenosti (distance) v jednotlivých atribútoch. V prípade ak 2 záznamy obsahujú výrazne podobné hodnoty vstupných atribútov, je veľmi pravdepodobné že aj ich cieľové atribúty budú mať podobné hodnoty (za predpokladu že vstupné atribúty sú relevantné). V prirodzených systémoch so spojitými veličinami sa veľmi často používajú prístupy, vyšetrujúce dopad zmeny vstupu na zmenu výstupu. Takéto prístupy využí-

¹ <http://www.machine-learning.martinsewell.com/ensembles/ensemble-learning.pdf>

vajúce diferencie boli použité aj pri modelovaní v rámci kauzálnej analýzy [3],[4]. Sledovanie nie len hodnôt, ale aj zmien hodnôt teda umožňuje spresnenie výsledného modelu. To súvisí aj s narábaním s hodnotami v procese tréovania. V procese tréovania sa modeluje závislosť medzi vstupom (vstupmi) a cieľovým atribútom. Avšak bežne používané spôsoby tréovania zvyčajne nezohľadňujú súčasne viaceré záznamy a už vôbec nie rozdiel medzi ich hodnotami. Je to však pochopiteľné, vzhľadom na fakt, že zohľadnenie takýchto rozdielov by bolo výrazne časovo náročné. Avšak, výnimku tvorí model k-najbližších susedov (ktorý do veľkej miery inšpiroval aj vznik tejto transformácie), ktorý síce model ako taký netrénuje, avšak zohľadňuje aj rozdiely hodnôt v atribútoch, z ktorých nakoniec počíta vzdialenosti záznamov.

V druhom rade sa jedná o štatistický fakt, keďže z väčšieho množstva nezávislých odhadov, dokážeme získať presnejšiu predpoveď cieľového atribútu. Väčší počet odhadov taktiež umožňuje použitie rozličných stratégií určenia finálnej predpovedanej hodnoty (aritmetický priemer, váhovaný priemer, odstránenie extrémov, výber najbližších záznamov, prípadne ich kombinácie).

V porovnaní s pôvodnou verziou prístupu [5] využívajúceho dátovú transformáciu, bolo vykonaných niekoľko zmien. V procese predikcie neboli použité všetky dostupné záznamy z tréovacej množiny (tak ako v pôvodnej verzii), ale len K záznamov s najnižšou euklidovou vzdialenosťou voči predikovanému záznamu. Hodnotu parametra K teda môžeme podľa potreby ladiť, pre dosiahnutie lepších výsledkov metódy; v našom prípade bola zvolená hodnota $K = 10$. Na vybrané záznamy bol aplikovaný natréňovaný regresný model, čím sme získali $2K$ odhadov cieľovej hodnoty. Ďalším rozdielom v porovnaní s predchádzajúcou verziou prístupu, je použitie váhovania pri priemerovaní získaných odhadov. Váhy jednotlivých záznamov boli určené na základe prevrátenej hodnoty vzdialenosti. Pre zabránenie deleniu nulou - v prípade ekvivalentných záznamov bola k vzdialenosti pripočítaná konštanta 0,01. Ďalším rozdielom, oproti pôvodnej verzii bolo použitie normalizácie vstupných atribútov dostupných údajov na interval 0 až 1. Dôvodom bolo zabránenie vplyvu rozdielnych rozsahov jednotlivých atribútov na určenie vzdialenosti dvojice záznamov.

Pre objektívnejšie zhodnotenie vhodnosti transformácie a výkonnosti modelov bola validácia vykonaná 10-násobne. Pri každom z 10 opakovaní, tréovacia množina pozostáva z náhodne vybraných záznamov spomedzi dostupných, pričom testovacia množina obsahovala zvyšné záznamy. Podmienky pri porovnávaní dosiahnutých výsledkov s použitím a bez použitia dátovej transformácie boli totožné (boli dokonca použité rovnaké seedy pre náhodný výber záznamov do tréovacej množiny). Taktiež z dôvodu vyššej objektívnosti boli použité 2 typy regresných modelov - neurónová sieť a strojový model M5P, ako aj viaceré dátové množiny. Pri experimentoch, bol okrem aspektu presnosti predikcie sledovaný aj časový aspekt predikcie, ako aj možnosť určenia intervalového odhadu cieľového atribútu pre konkrétny predikovaný záznam.

2 Dosiahnuté výsledky

Cieľom tohto článku je otestovať prezentovanú dátovú transformáciu s modifikovanou stratégiou určenia finálnej hodnoty na reálnych dátových množinách. Ako dátové množiny boli použité množiny Combined Cycle Power Plant Data Set² (označená ako PowerPlant) a Energy Efficiency³. Uvedené množiny obsahujú 9568 a 768 záznamov pri 4, resp. 8 číselných atribútoch. Pre obe dátové množiny boli realizované rovnaké experimenty. Z dátovej množiny bolo náhodne zvolených 200 záznamov, ktoré tvorili trénovaciu množinu.

V prvej fáze boli natrénované 2 typy regresných modelov (neurónová sieť a regresný strom) nad originálnymi dátami. Každé trénovanie bolo realizované 10 krát, s rozdielnymi hodnotami seedu, pre odlišné inicializačné nastavenie siete, ako aj odlišne zvolené záznamy v trénovacej množine. Na zvyšných záznamoch boli modeli validované, pričom z 10 opakovaní bol určený priemer. Celý tento proces bol opakovaný aj pre nižší počet záznamov - 195, 190, 185, ... 80. V druhej fáze boli za rovnakých podmienok (pri použití rovnakých hodnôt seedu, ako aj rovnakých vybraných záznamov) vytvorené aj modeli z transformovaných dát.

Tabuľky Tab 1. a Tab 2. demonštrujú priemerné dosiahnuté presnosti modelov vyjadrené korelačným koeficientom (KK) a strednou kvadratickou chybou (SKCH). Priemerné hodnoty sú vypočítané vždy z 10 realizovaných opakovaní. Ako prvý model bola použitá viacvrstvá neurónová sieť perceptronov s 2 skrytými vrstvami, pričom aktivačnou funkciou bol sigmoid. Koeficient učenia bol zvolený 0.3 a maximálny počet epoch bol nastavený na 500. Druhým modelom bol regresný strom M5P [7] pri použití minimálneho počtu záznamov na list 4, s orezaním.

Tab 1. Porovnanie výkonnosti natrénovaných regresných modelov s použitím a bez použitia dátovej transformácie na dátovej množine PowerPlant.

| | | Model neurónová sieť | | Stromový model M5P | |
|-----------------|------|----------------------|-----------------|--------------------|-----------------|
| | | 140 záznamov | 180 záznamov | 140 záznamov | 180 záznamov |
| s transform. | KK | 0.9595 | 0.9646 | 0.9652 | 0.9629 |
| | SKCH | 4.8021 | 4.5278 | 4.4963 | 4.6251 |
| bez transf. | KK | 0.9656 | 0.9663 | 0.9636 | 0.9641 |
| | SKCH | 5.1683 | 5.0259 | 4.5938 | 4.5534 |

² ML Repository: <http://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant>

³ Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets/Energy+efficiency>

Tab 2. Porovnanie výkonnosti natrénovaných regresných modelov s použitím a bez použitia dátovej transformácie na dátovej množine Energy Efficiency.

| | | Model neurónová sieť | | Stromový model M5P | |
|-----------------|-----|----------------------|-----------------|--------------------|-----------------|
| | | 140 záznamov | 180 záznamov | 140 záznamov | 180 záznamov |
| s transform. | KK | 0.9977 | 0.9993 | 0.9982 | 0.9993 |
| | KCH | 0.8642 | 0.6981 | 0.8308 | 0.7035 |
| bez transf. | KK | 0.9776 | 0.9862 | 0.9804 | 0.9856 |
| | KCH | 2.2451 | 1.9694 | 2.1177 | 1.8697 |

Trénovacie množiny pozostávali zo 140 resp. 180 náhodne vybraných záznamov, testovacie množiny obsahovali zvyšné nepoužité záznamy. Ako modeli boli použité viacvrstvové neurónové siete perceptronov, ako aj regresné stromy M5P. Pri trénovaní modelov bola použitá knižnica Weka. Pri validácii boli vyčíslené kritériá korelačný koeficient (KK) a stredná kvadratická chyba (SKCH). Výsledné výkonnosti uvedené v tabuľkách predstavujú priemer z 10 validácií.

Z dosiahnutých výsledkov v Tab. 2 si môžeme všimnúť pri použití transformácie výrazný nárast presnosti natrénovaných modelov v oboch sledovaných kritériách. V tabuľke sú uvedené iba prípady s počtom záznamov 140 a 180, avšak aj ostatné testované prípady s počtom záznamov 80 až 200 dosahujú výrazne podobné výsledky.

Pri testovaní transformácie na dátovej množine PowerPlant, ktorej výsledky sú uvedené v Tab. 1, spresnenie nie je až natoľko výrazné. Obzvlášť je to zrejme v prípade použitia regresných stromov M5P. Pri použití neurónových sietí sa so zvyšujúcim počtom záznamov v trénovacej množine zvyšuje aj presnosť modelu. Celkovo však modeli natrénované za použitia dátovej transformácie dosahujú v priemere lepšiu výkonnosť, len v niekoľkých individuálnych prípadoch dosiahli o niečo horšiu výkonnosť.

Z časového aspektu, je trénovanie a predikcia za použitia prezentovanej dátovej transformácie výrazne časovo náročnejšia. Tento aspekt je však očakávaný, vzhľadom na potrebu viacnásobného aplikovania modelu ako aj dátovej transformácie na predikované dáta. Predikcia za použitia transformácie je približne stonásobne pomalšia, v závislosti od typu stratégie určenia finálnej predikovanej hodnoty a počtu odhadov. Prezentovanú techniku je preto vhodné použiť v prípade, ak primárnym kritériom je vysoká presnosť modelu a prípadné vyššie časové nároky nie sú prekážkou.

3 Záver

Celkovo, prezentovaná transformačná technika vykazuje potenciál, spočívajúci v zvýšení presnosti regresných modelov. Ukázalo sa to na syntetických [5] ako aj reálnych dátach, pričom zlepšenie presnosti modelu bolo zrejme z oboch sledovaných kritérií - korelačného koeficientu, ako aj strednej kvadratickej chyby. Z časového hľadiska, použitie tejto techniky značne zvyšuje časovú náročnosť (obzvlášť vo fáze predikcie), čo je však efekt, ktorý bol pri návrhu techniky očakávaný. Je preto vhodné zvážiť použitie tejto techniky v závislosti od požiadaviek na presnosť a rýchlosť predikcie modelu, ako aj počtu záznamov v trénovacej množine. Celkovo však prezentovaná transformačná

technika vykazuje viacero pozitívnych aspektov, medzi ktoré patria aj možnosť voľby typu modelu, možnosť realizácie intervalového odhadu cieľovej hodnoty, možnosť voľby stratégie určujúcej výpočet cieľovej hodnoty ako aj výrazné spresnenie modelov. Je zrejmé, že nie u všetkých reálnych dátových množín dôjde k takto výraznému zvýšeniu presnosti. Do budúcnosti tak zostáva potreba podrobnejšie otestovať techniku na ďalších dátových množinách. Zaujímavé by tiež bolo porovnanie presnosti modelov používajúcich prezentovanú techniku a metódu boostingu.

Pod'akovanie: Táto publikácia vznikla vďaka podpore projektu VEGA 2/0167/16.

Literatúra

1. Dietterich Thomas G.: Ensemble Methods in Machine Learning, Oregon USA, 1998.
2. Elith Jane, Leathwick John: Boosted Regression Trees for ecological modeling, 2016.
3. Kvassay M., Hluchý L., Krammer P., Schneider B.: Causal analysis of the emergent behavior of a hybrid dynamical system. In Acta polytechnica Hungarica: journal of applied sciences at Budapest Tech Hungary, 2014, vol. 11, no. 4, p. 21-40. (0.471 - IF2013). ISSN 1785-8860.
4. Kvassay M., Krammer P. Hluchý L., Schneider B.: Causal Analysis of an Agent-Based model of Human Behaviour, Computing and Informatics, 2016, vol. 32. (in review)
5. Krammer P., Hluchý L.: Transformačná regresná technika pre dolovanie v údajoch. WIKT 2014: 9th Workshop on Intelligent and Knowledge Oriented Technologies, Bratislava, p. 45-50, ISBN 978-80-227-4267-2.
6. Schonlau Matthias R.: Boosted Regression (Boosting), The Stata Journal 5, Number 3, 2005, pp. 330 - 3654.
7. Wang Y., Witten I. H.: Induction of model trees for predicting continuous classes, In Poster papers of the 9th European Conference on Machine Learning, 1997.

Annotation:

Using Transformation Regression Technique for Data Mining

Data mining and machine learning methods can be used in many domains. However, several domains generate limited volume of data only, because getting larger data sets is difficult from time, economical, or technical aspects. But these domains also require a modelling and predicting; so the small data volume can cause problems in generalization and decreasing of model precision. Presented paper deals about data transformation, which original regression task replace with regression task, with higher count of records, with tendency to increase model precision. Paper demonstrate a new modification of data transformation which testing on real data sets. Reached performance comparison and evaluation are published in paper.

Strojové učení pro analýzu rodinného podnikání

Juraj Michalik, Luboš Popelínský, Klára Antlová a Petra Rydvalová

KD Lab, FI MU Brno a FE TU Liberec

popel@fi.muni.cz

Abstrakt. Po stručném úvodu do problematiky rodinného podnikání popíšeme, jaká data jsou veřejně dostupná, a naznačíme možnosti metod strojového učení pro analýzu dat souvisejících s rodinným podnikáním. Uvádíme též první výsledky z pilotní studie o 67 rodinných podnicích.

Typ příspěvku: Příspěvek o probíhajícím výzkumu

Klíčové slová: rodinné podnikání, strojové učení, klasifikace

1 Úvod

Rodinné podnikání je zajímavou alternativou k masové anonymizované produkci z netvůrčí práce a v řadě nám blízkých společenstvech je podporováno státem. Podobně je tomu i v České republice. Poněkud opožděný je však nejen rozvoj rodinného podnikání u nás, ale i legislativní proces. Přitom podle Karla Havlíčka, předsedy představenstva Asociace malých a středních podniků a živnostníků ČR (AMSP) je v ČR 260 000 malých a středních podniků, z nichž cca 70% může být podniků rodinných, dle odhadu AMSP na vzorku (viz konference *Budoucnost rodinných firem v ČR - inspirace i výzvy*, 24. února 2016, uspořádaná u příležitosti spuštění webového portálu o rodinných firmách *majitelefirem.cz*.) Podle Jiřího Hnilici (VŠE, tamtéž) jednotná definice neexistuje. Jednotlivé definice však obvykle za zásadní považují vlastnický podíl rodiny, především

1. existující záměr firmu předat v rámci rodiny,
2. směřování ovlivňuje více než jedna generace a
3. rodina nejenom firmu vlastní, ale podílí se na řízení.

Více k této problematice u nás viz [4].

Každá země zejména na západ od Smolenic, ale i např. Slovensko, specifikaci rodinného podniku v zákoně ukotvenou má. Bod 1. asi sotva můžeme zjistit jinak než přímým dotazem. Pro tento příspěvek považujeme za **rodinný podnik** takový, který splňuje bod 3. a slabší variantu bodu 2. - ve vedení firmy jsou alespoň dva členové rodiny. Navíc požadujeme, aby majetkový vklad rodiny byl rozhodující.

V projektu *TACR Rodinný podnik – řešení sociálních a ekonomických disparit obcí*, jehož je tato práce součástí, se pokoušíme metodami analýzy dat přispět k popisu existujícího stavu a k vyjasnění některých pojmů (či postupů) s rodinným podnikáním spojených. Jedním z cílů může být zjištění, jak pro různé definice rodinného podniku jsem schopni takový rodinný podnik rozpoznat automaticky a jak jsou existující data vhodná pro další analýzu, např. shlukování, predikci nebo detekci anomálií, metodami strojového učení [2].

Předpokládáme, že popis stavu rodinného podnikání v naší zemi a určení vitality (zdraví) rodinného podniku jsou dvě oblasti projektu, kde strojové učení může být užitečné. Tento text nám slouží též k prvnímu zamýšlení nad takovým použitím.

2 Informace o rodinném podnikání

2.1 Vyhledávače a portály

První otázkou bylo, zda je možné rodinné podniky najít pomocí webového vyhledávače. Zkoušeli jsme tři jednoduché dotazy (použit Maxthon, svobodný multiplatformní šestý celosvětově nejpoužívanější webový prohlížeč vyvíjený společností Maxthon International) a sledovali pro prvních 20 odkazů přesnost (precision, tj. kolik z vyhledaných odkazů splňuje naši definici)

| | | |
|--------------------------------|-------|-----|
| <i>a syn firma</i> | 16/20 | 80% |
| <i>rodinná firma s tradicí</i> | 13/20 | 65% |
| <i>tradiční rodinná firma</i> | 10/20 | 50% |

I když tento experiment nebyl rozsáhlý, vidíme, že deklarovaný rodinný podnik nemusí rodinným podnikem být.

Takto nalezené stránky jsou ovšem vytvořeny pro jiný účel, a proto téměř neobsahují důležité ekonomické informace. Pro to jsou vhodnější data z portálů www.detail.cz, případně www.Merk.cz a <http://www.info.mfcr.cz/ares/>. První z nich podává o každé z firem velmi dobrý přehled. Tato placená služba totiž poskytuje souhrn všeho, co se dá na internetu o firmě najít. Pro naše potřeby se zdá dostačující třetí, *ARES - Administrativní registr ekonomických subjektů*, i když v něm některá data chybí. Obsahuje kromě adresy a hlavní oblasti podnikání mj. i údaje z veřejné části Živnostenského rejstříku včetně historie, a informaci o spolehlivosti plátce DPH.

2.2 Data z dotazníkového šetření

Další možností, jak získat přesnější data, jsou různé typy dotazníkových šetření. První pilotní dotazníkové šetření nedávno provedené Technickou univerzitou Liberec [3] (více viz <http://vyzkum.ef.tul.cz/td03000035/>) obsahuje údaje o 67 rodinných podnicích. Protože se dat týká i následující analýza, uvádíme seznam jednotlivých položek, z nichž se záznam o rodinném podniku skládá, a jejich typu. Jedná se o Název

firmy, IČO, Typ podnikání (SRO,AS,FO,VOS), Datum, kdy společnost začala existovat, Jak je společnost stará, Počet zaměstnanců, Kód oblasti podnikání, Zda firma čerpala státní účelové dotace, Objem čerpaných dotací, Firma začala jako živnostník, Insolventnost, Věk zakladatele, Kapitál firmy, Je plátcem DPH, Kraj (LB,USTI,PR,ZLIN,CB) a Smlouva s obcí.

3 Experimenty s daty z dotazníkového šetření

Protože jsme v době odevzdání tohoto textu neměli k dispozici data o ne-rodinných podnicích, omezujeme se tu jen na rozpoznání některých charakteristik rodinných podniků. Použili jsme vždy všechny atributy popsané nahoře kromě IČO. Jako kritérium kvality jsme pro klasifikaci zvolili celkovou správnost (accuracy – relativní počet správně klasifikovaných instancí z testovací množiny) a pro regresní úlohy korelaci a RRSE (relative root squared error v %). Hodnota baseline odpovídá danému kritériu při náhodné klasifikaci. Pro analýzy jsem využili nástroj Weka [1] a všechny metody s defaultním nastavením, 10ti složkovou křížovou validaci. Uvádíme výsledky jen pro ty atributy, kde byl rozdíl oproti baseline výrazný (více než 5% u klasifikačních úloh). Vždy je uvedena učící metoda s nejlepší přesností a baseline.

Je plátcem DPH. Pro rozhodovací strom (metoda J48) obsahující pouze tři atributy - Capital, Employee, Birth of pioneer - celková správnost dosáhla 88.1% při baseline 71.7%. Poměrně silná závislost.

Firma začala jako živnostník. Nejlepší výsledek jsme dosáhli s Random Forest, kde správnost přesáhla 67.2 %, baseline 56.7%. Závislost tohoto atributu na zbývajících je tedy slabá.

CEDR - byla příjemcem dotací. Po odstranění atributu s objemem dotací byla správnost 79.1% (J48, baseline 58.2 %) a fakt získání dotací závisel jen na počtu zaměstnanců – čím větší firma, tím větší pravděpodobnost získání dotace.

Věk zakladatele. Pro tuto regresní úlohu jsme použili Random Forest. Korelační koeficient přesáhl 0.57 (baseline -0.35), RRSE 89.9% .

Počet zaměstnanců. Druhou regresní úlohou bylo zjištění, zda je počet zaměstnanců závislý na ostatních attributech. Zde korelační koeficient dosáhl 0.45 a RRSE 93.6 %. Závislost tedy existuje, ale není silná.

4 Závěr

V této krátké úvodní studii jsme se soustředili na rodinné podnikání z pohledu zpracování dat a uvedli první výsledky ze zpracování pilotní studie o rodinných podnicích. V současné době (říjen 2016) byl ukončen sběr dat, která popisují stav rodinného podnikání v malých sídlech na celém území Česka. Na jejich základě sestavujeme datovou sadu negativních příkladů, tj. ne-rodinných podniků, které se rodinným podnikům co nejvíce podobají, či dokonce rodinnost deklarují a přitom rodinnými nejsou.

Poděkování. Děkujeme Pavle Suchánkové a Kateřině Bekové za úvodní experimenty. Tato práce byla částečně podporována grantem TAČR - Omega, TD03000035 Rodinný

podnik – řešení sociálních a ekonomických disparit obcí a Fakultou informatiky Masarykovy university Brno.

Literatura

1. Aggarwal, C.C.: Outlier Analysis. Springer, (2013). Friedman, A.D., Menon, P.R.: Theory and Design of Switching Circuits. Computer Science Press, Inc., (1975).
2. Hall m. et al.: The Weka data mining software. An update. SIGKDD Exploration Newsl., 11(1):10{18, November 2009.
3. Rydvalová P., Karhanová Horynová E., Jáč I., Valentová E., Zbránková M.: Rodinné podnikání – zdroj rozvoje obcí. 1. vyd. Liberec: Technická univerzita v Liberci (2015).
4. Rydvalová, P., Karhanová Horynová E., Zbránková M.: Family Business as Source of Municipality Development in the Czech Republic. Amfiteatru Economic. 1st ed. Bucharest: The Bucharest Academy of Economic Studies, vol. 18, iss. 41, pp. 168 - 183 (2016)

Annotation:

Machine learning for family business analysis

We give a brief overview of family business, mention difficulties when formulating a definition of a family enterprise and then describe what data are nowadays available for data analysis. We show how machine learning methods can be used for analysis of a 67 family enterprises. In conclusion we mention future work.

Anomaly detection for aircraft engine fault prediction

Tomáš Rudolecký

KD Lab, FI MU Brno

454906@mail.muni.cz

Abstract. Aircraft engine failures can be expensive and an obvious security threat. When we are able to predict a potential failure of an engine in advance, then we can send the aircraft for maintenance. Sensor data is collected during engine starts, takeoffs, cruise or special events. Aim of this research is to create a model of standard behavior of so called healthy engines and based on that, detect serious change which can predicts a failure. Furthermore, we want to distinguish among particular failure types. The model don't have just to be able to successfully pass data tests but also should have some physical explanation. Sometimes the resultant model shows big dependences on attributes which should be at most auxiliary, or it shows physically improbable relations among attributes. We present the first results obtained with One-class Support Vector Machine, which show significant increase of the anomaly factor of two out of four faulted engines when they were approaching the failure. We also made experiments with group anomaly detection.

Contribution type: Work-in-progress paper

Key words: fault prediction, aircraft engine, support vector machine, group anomaly detection

1 Introduction

Every flight phase is getting different sensor records which are captured as one record per event. It can be one snapshot at the time when the event occurred like takeoff, or snapshot of sensor records captured during some time period like engine start up.

In this paper we are focused on engine starts. During a start, an engine goes through a number of phases during which various components become dominant. These components are measured as fuel flow, speed of high pressure compressor, times between phases, various temperatures and pressures. The start phases are related to a different types of a failure.

Our data contains over a million of records from over a thousand engines. We have a list of engines built on maintenance reports for every type of fault. The predictive model should recognize increasing probability of a fault with no false negatives and reasonable number of false positives.

Common predictive algorithms don't show any difference between the classes and anomaly detection applied on a non-transformed data shows roughly the same ratio of anomalies for healthy and faulted engines. Our research seems challenging, if we take into account the fact, that most of predictive algorithms for aircraft engines are based just on few attributes with decision boundaries defined by experts.

2 Feature engineering

2.1 Domain knowledge

Domain knowledge is necessary in the phase of attribute selection and feature extraction. Engine experts identified list of attributes which may be related to a certain fault, thus we have subset of reasonable attributes to begin with. When we encounter a false positive engine, we can ask whether the engine didn't have some unusual maintenance or other condition which would eliminate it from the testing set of healthy engines.

2.2 Data transformation

Feature engineering aims to describe inherent structures which would create the best representation of the data.

Most of the attributes are affected by a seasonal effect which is easily visible on plots as a periodical wave. Therefore, we use polynomial regression with ambient temperature as the independent variable, to eliminate that.

Values usually oscillate, thus we use moving averages to capture trends. We also derived differences between successive records and moving variances. Principal components were added to represent linear interactions among features.

3 One class Support vector machine for anomaly detection

Support vector machine projects data through a non-linear function to a more dimensional space in which then separates the data to classes. Other possibility is to use a kernel function to create a non-linear boundaries without projection to a new feature space.

One class SVM performs unsupervised learning. Training samples define a function that takes the value +1 in a small region capturing most of the data points and -1 elsewhere [1]. The class of training samples is separated by hyperplane with maximal distance from origin. Normal data provided to SVM creates a representational model. New data presented to the model is then assessed by probability of being inside of the model.

Aggregation of anomaly points can represent each engine and we can simply sort engines according to their ratio of those points. Nevertheless, the problem with such approach is that it is hard to decide how many records should be included in one group representing particular engine. In other words, the question is how many days/records before the fault we should see signs of it? Another issue is, that if the ratio is high but decreasing before the fault, it doesn't indicate a coming fault.

We have tried to create the model based on healthy engines, which define normal behavior and indicate any significant change in it. Other option is to create the model based on records from engines with a particular failure, and thus separate small class of records. The second approach generates almost no false positives, but when we remove one faulty engine from training set and then use the engine in testing, it is recognized as healthy for most of the models. Thus, the model is overfitting the training data.

4 Group anomaly detection

Common anomaly detection looks at every record individually, but we can have records which aren't anomalies themselves, but their distribution as a group is different from other groups. This approach was successfully used for astronomical data and for data from high energy particle physics experiments [2].

For our experiments we chose Mixture of Gaussian mixture model (MGMM) which works with multimodal distributions. MGMM assume, that feature vectors in groups, are generated by a mixture of K Gaussian distributions [3].

5 Results

Successful model should be able to recognize increasing ratio of anomalies for faulted engines. We found this behavior in two out of four faulted engines with fuel system failures. Results of one class SVN are depicted on Figure 1, where the anomaly score is the anomaly level acquired from the algorithm, when the required percentage of anomaly points is set to 1%. Healthy engines have anomalies, but none with such an increase. Anomalies in healthy engines and faulted engines outside the timeframe before the fault, have intermittent character. In other words, they don't have many consecutive anomalies. Now we are investigating other types of start-up failures. Interesting is, that when we took all types of failures together to see, whether we can find anomalies indicating any type of problem, we got many false positives and false negatives (depending on algorithm setting), and thus we conclude, that they don't have common model.

Group anomaly detection with MGMM didn't show any difference between healthy and faulted engines, even with various settings of parameters, combinations of features and number of records per engine.

6 Discussion

Our research leads us to time series which logically seems as an appropriate solution. Nevertheless, we have to deal with the fact of irregularity of data capturing. We can have many records in one day or few weeks without any record. Records which happened in one day can be highly dependent on each other. Missing data can be caused by engine inactivity or just the fact we didn't get the data. These conditions should be included in the predictive algorithm.

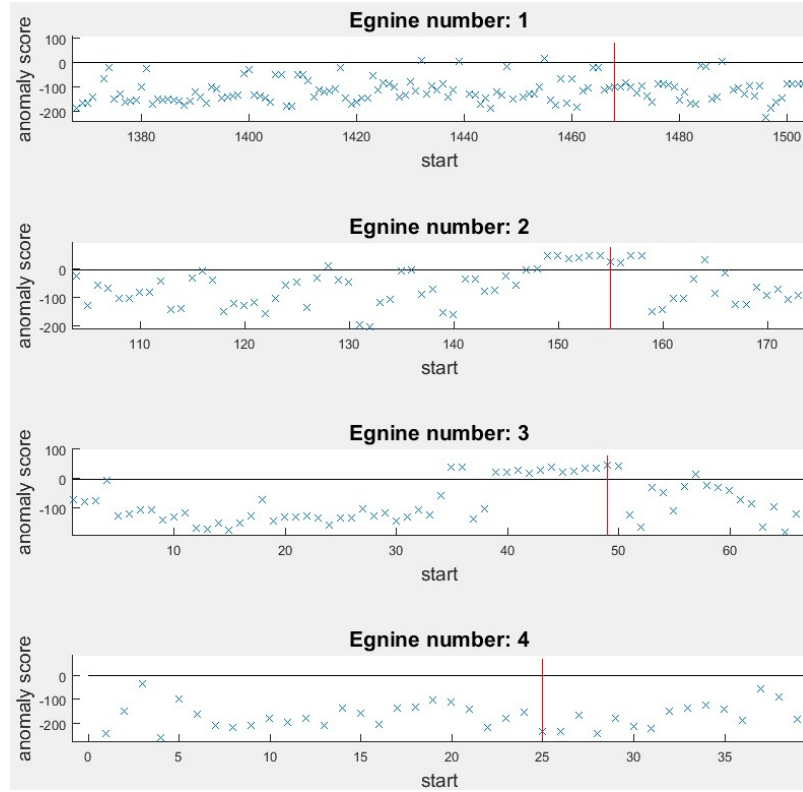


Fig. 3. Results from one class SVM show increasing anomaly level of second and third engine before the fault occurred. Red vertical line indicates time of the fault and the black horizontal line separates normal records (below the line) and anomalies (above the line).

References

1. Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., Platt, J.: Support Vector Method for Novelty Detection. MIT Press. 1999, 2000(12), 582-588.
2. Muandet, K., Schölkopf, B.: One-Class Support Measure Machines for Group Anomaly Detection. Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence. 2013, 2013(6).
3. Xiong, L., Póczos, B., Schneider, J., Connolly, A., Vanderplas, J.: Hierarchical probabilistic models for group anomaly detection. JMLR WCP Proceedings of the International Conference on Artificial Intelligence and Statistics AISTATS. 2011, 2011(15), 789-797.

Systém na podporu rozhodovania pomocou jednoduchého a efektívneho pochopenie medicínskych záznamov

Michal Vadovský, František Babič, Miroslava Muchová

Katedra kybernetiky a umelej inteligencie
Fakulta elektrotechniky a informatiky
Technická univerzita v Košiciach
Letná 9/B, 042 00, Košice, Slovenská republika

{michal.vadovsky, frantisek.babic, miroslava.muchova}@tuke.sk

Abstrakt. Medicínske záznamy predstavujú dôležitý zdroj informácií o zdravotnom stave pacientov, ale často sú uchovávané vo forme, ktorá neumožňuje ich efektívnu správu a najmä využitie pre účely medicínskej diagnostiky. Cieľom našej práce bolo ukázať potenciál vybraných metód exploračnej analýzy a dolovania v dátach práve na jednoduché a efektívne pochopenie medicínskych záznamov. Na tento účel sme použili vzorku dát z Chorvátska, v rámci ktorej sú jednotliví pacienti charakterizovaní širokou škálou parametrov, bežne zisťovaných a vyhodnocovaných v ambulanciách praktického lekára. Na základe týchto dát sme sa snažili identifikovať kľúčové symptómy alebo hraničné hodnoty pre diagnostiku ochorenia s názvom Mierne kognitívne zhoršenie. Dosiahnuté výsledky sú prezentované v jednoducho pochopiteľnej forme aj pre používateľa s minimálnymi znalosťami zo štatistiky alebo analýzy dát – lekára.

Typ príspevku: Výskumný príspevok

Kľúčové slová: dáta o pacientoch, analýza, diagnostika

1 Úvod

Mierne kognitívne zhoršenie (v angl. „Mild Cognitive Impairment“ (MCI)) predstavuje medzistupeň medzi očakávaným úbytkom kognitívnych funkcií normálneho starnutia a vážnejším nástupom demencie [1]. Problémy s pamäťou, rozprávaním a myslením nastupujú rýchlejšie ako je obvykle pri bežnom starnutí obyvateľov. Výskyt MCI môže zvýšiť riziko neskoršieho vývoja demencie spôsobenej Alzheimerovou chorobou (ACH) alebo inými neurologickými poruchami. Na druhej strane, zdravotný stav ľudí s touto poruchou sa časom môže aj zlepšiť. Medzi rizikové faktory s najväčším vplyvom na pozitívnu MCI diagnostiku patria napríklad rastúci vek alebo špecifická forma génu známeho ako APOE-e4, ktorý je tiež spájaný s ACH. Taktiež sa zohľadňuje ži-

votný štýl jedinca, ale dôkazy o týchto rizikových faktoroch už nie sú také jasné (cukrovka, fajčenie, depresie, vysoký krvný tlak, zvýšený cholesterol, nedostatok telesného pohybu). Včasná diagnostika tohto ochorenia umožní okamžite nasadiť vhodné liečebné postupy, pomocou ktorých je možné napríklad spomaliť alebo zabrániť progresívnej strate pamäti [4]. Bežný postup pri takejto diagnostike je postupný zber všetkých potrebných vstupných dát, na základe ktorých si lekár následne vytvorí celkový obraz o zdravotnom stave pacienta a urobí rozhodnutie. Tento postup je však vo väčšine prípadov pomerne časovo náročný a najmä si vyžaduje neustály prehľad a pochopenie stále rastúceho objemu dát. To otvára priestor pre vznik a nasadenie systému na podporu rozhodovania, pomocou ktorého bude môcť lekár spracovať a pochopiť nielen údaje o aktuálnom zdravotnom stave pacienta, ale konfrontovať ich aj s historickými hodnotami.

Vzorka dát obsahuje informácie o 93 pacientoch z klinickej praxe, ktorú vykonáva spolupracujúci expert v Chorvátsku. Medzi týmito pacientmi sa nachádza 35 mužov a 58 žien vo vekovom intervale 50 až 89 rokov, u ktorých je pomer pozitívna vs. negatívna diagnostika MCI 37 ku 56. Každý pacient je zároveň charakterizovaný hodnotami 59 faktorov, ktoré predstavujú potenciálne dôležité vstupy pre diagnostiku MCI, napr. vek, pohlavie, hypertenzia, cukrovka, cholesterol, kardiovaskulárne ochorenia, alergie, úroveň bielych krviniek, úroveň červených krviniek, atď. Cieľová diagnostika v našom prípade predstavuje binárny atribút: (0) – zdravý človek, (1) – pozitívne diagnostikované ochorenie MCI.

Článok je rozdelený na niekoľko hlavných častí. Úvod je venovaný predstaveniu problému, použitých metód a dátovej vzorky. Druhá časť je venovaná návrhu systému s cieľom poskytnúť lekárovi dôležité informácie pre podporu jeho rozhodovania. Záver sumarizuje dosiahnuté výsledky a načrtáva ďalšie kroky autorov v tejto problematike. Zároveň je potrebné spomenúť, že autori túto dátovú sadu alebo jej podobné analyzovali aj pomocou iných metód dolovania v dátach, napr. vybraných algoritmov pre generovanie rozhodovacích stromov. Výsledky týchto experimentov sú prezentované v článkoch [2, 3]. V tomto článku bolo hlavnou motiváciou overiť potenciál vybraných štatistických metód pre prezentáciu dôležitých zistení extrahovaných z dát v jednoducho pochopiteľnej forme koncovým používateľom, t.j. lekárom. Jazyk R na tento účel ponúka viacero možností, ktorých výsledkom môže byť jednoducho ovládaný systém na podporu rozhodovania.

2 Návrh systému na podporu rozhodovania

Prototyp navrhovaného systému je výpočtovo založený na metódach popísaných vyššie, ktoré sú implementované v jazyku R. Používateľské rozhranie je implementované prostredníctvom aplikačného balíka RShiny. Platnosť dosiahnutých výsledkov je možné overiť prostredníctvom existujúcej medicínskej literatúry alebo na základe praktických skúseností spolupracujúceho experta.

2.1 Identifikácia kľúčových faktorov

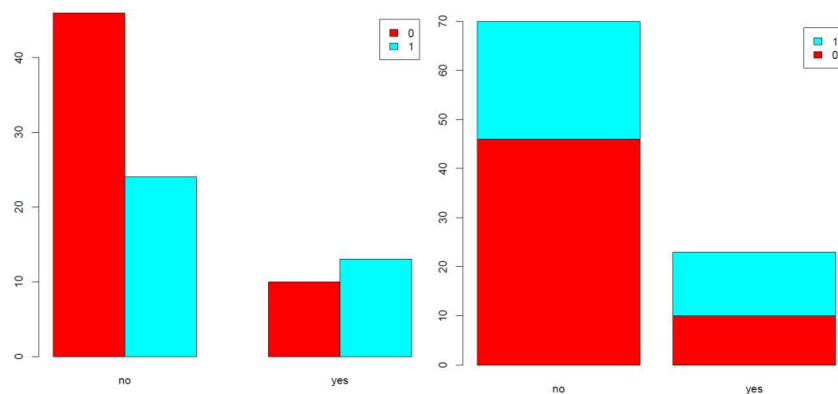
Na analýzu závislostí medzi cieľovým a vstupnými atribútmi sme sa rozhodli použiť 2-výberový Welchov t-test pre numerické atribúty [6] a Pearsonov Chi-kvadrát test nezávislosti pre nominálne atribúty [6]. V prvom prípade sme si stanovili alternatívnu hypotézu H_A , ktorá vyjadruje, že rozdiely medzi priermi populácie (0/1) sú rôzne (závislosť atribútov). K nej nultá hypotéza (H_0) vyjadruje, že priemer populácie je rovnaký (nezávislosť atribútov). Najvyššie závislosti sme identifikovali pre nasledovné numerické atribúty: Vek ($p\text{-value}=0.000017$, hladina významnosti 0.01), Alpha-2 globulín (0.0458, 0.05), (0.0535, 0.1) a Skinf (0.0953, 0.1). Atribút Clear predstavuje dobrú charakteristiku filtračnej kapacity obličiek; nízka alebo znížená hodnota znamená chronické ochorenie obličiek. Atribút Skinf definuje hrúbku kožnej riasy na tripepe. Podobný postup sme použili aj pre nominálne atribúty; nulová hypotéza v tomto prípade tvrdí, že medzi dvoma nominálnymi atribútmi nie je závislosť. V množine 17 atribútov sme potvrdili zaujímavú závislosť na hladine významnosti 0.1, t.j. na 90%, len v prípade Analg (liečba analgetikami). Túto závislosť potvrdzuje aj Obr.1.

2.2 Identifikácia hraničných hodnôt

Hraničné hodnoty pre jednotlivé vstupné atribúty sme analyzovali pomocou ROC krivky a výpočet zlomového bodu Youden metódou [5]. Táto metóda je bežne používaná na vyhodnotenie testov v bioštatistike [7].

$$J(c) = \max \{Sensitivity(c) + Specificity(c) - 1\} \quad (1)$$

J – je funkciou návratnosti (senzitivity) a špecificity; v optimálnej hodnote c je maximálna vertikálna vzdialenosť medzi ROC krivkou a hlavnou diagonálou

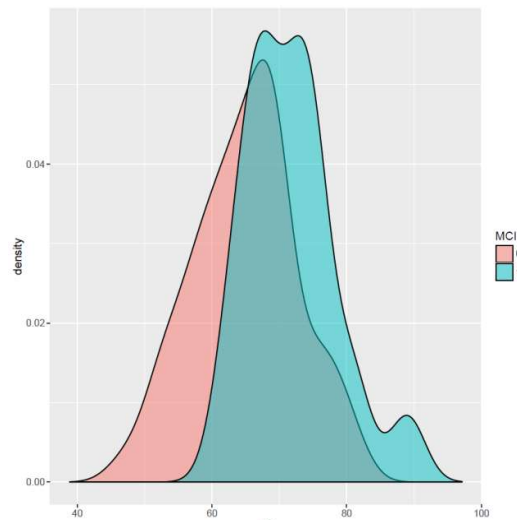


Obr. 1. Vizualizácia závislosti medzi cieľovou diagnostikou MCI (y, 0/1) a vstupným atribútom Analg (x, yes/no)

Na Obr.2 je zobrazený graf distribučnej funkcie, ktorá nám ukazuje, že pacienti s pozitívnou MCI diagnózou sú vo vyššom veku, ako v prípade zdravých ľudí. Zlomový bod

v tomto prípade predstavuje vek 70.5 rokov, v rámci ktorého bola dosiahnutá presnosť klasifikácie 70.97%.

Používateľ má zároveň k dispozícii aj ďalšie metriky na vyhodnotenie úspešnosti určenia zlomového bodu, okrem presnosti. Napr. parameter v angl. „true positive“, t.j. koľko pozitívnych príkladov bolo v skutočnosti klasifikovaných správne. Nízka hodnota tohto parametra znamená, že viacero negatívnych prípadov bolo označených ako pozitívne, čím generovali tzv. falošný alarm. V tomto prípade je potrebné vziať do úvahy náklady na potrebnú liečbu.



Obr. 2. Vizualizácia distribučnej funkcie pre atribút Vek (x) a cieľovou diagnostikou MCI (y, 0/1) s identifikovaným zlomovým bodom

3 Záver

Cieľom článku bolo priblížiť možnosti exploračnej analýzy a dolovania v dátach pre jednoduché a efektívne pochopenie medicínskych záznamov. Tieto záznamy predstavujú dôležitý zdroj informácií, pomocou ktorých diagnostikuje lekár príslušné ochorenia. Často je potrebné zvážiť viacero symptómov meniacich sa v čase a v rôznych súvislostiach. To otvára priestor pre vytvorenie komplexného systému na podporu rozhodovania, ktorého odporúčania budú k dispozícii v graficky príjemnej a jednoducho pochopiteľnej forme. Príklady prezentované vyššie predstavujú len ukážku; rovnaké vizualizácie a popisné charakteristiky je možné generovať pre všetky vstupné atribúty. Viac detailov o experimentoch a navrhnutom systéme obsahuje článok podaný do časopisu „BMC Medical Informatics and Decision Making“.

Podakovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektu č.1/0493/16 financovaného Vedeckou grantovou agentúrou MŠVVaŠ SR a SAV (VEGA), projektu

č. 025TUKE-4/2015 financovaného Kultúrnou a edukačnou grantovou agentúrou MŠVVaŠ SR (KEGA), a interného výskumného projektu Fakulty elektrotechniky a informatiky č. FEI-2015-2.

Literatúra

1. Albert, M. S., a kol.: The Diagnosis of Mild Cognitive Impairment due to Alzheimer's Disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease. *Alzheimer's & Dementia* (2011), 7 (3), str. 270–79.
2. Babič, F. a kol.: On Patient's Characteristics Extraction for Metabolic Syndrome Diagnosis: Predictive modelling based on Machine Learning. *LNCS* (2014), Vol. 8649, str. 118–132.
3. Lukáčová, A. a kol.: How to Increase the Effectiveness of the Hepatitis Diagnostics by Means of Appropriate Machine Learning Methods. *LNCS* (2015), Vol. 9267, str. 81–94.
4. Petersen, R. C., a kol.: Mild Cognitive Impairment: Clinical Characterization and Outcome. *Archives of Neurology*, (1999) 56 (3), str. 303–308.
5. Schisterman, E.F., Perkins, N.J., Liu, A., Bondell, H.: Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology*, (2005), 16, str. 73–81.
6. Shahbaba, B.: *Biostatistics with R: An Introduction to Statistics Through Biological Data*. Springer, (2012).
7. Smith, N.: A note on Youden's J and its cost ratio. *BMC Medical Research Methodology* (2010).

Annotation:

Decision Support System based on simple and effective understanding of the available medical records.

The medical records are an important source of information about patient's health status, but they are often kept in a form which does not allow their effective management and their use for the purpose of the medical diagnosis. The aim of this work was to demonstrate a potential of the selected methods of the exploratory data analysis and data mining for a simple and effective understanding of the available medical records. For this purpose, we used a data sample from the Croatia, in which individual patients are characterised by a wide range of the parameters, routinely collected and evaluated by the general practitioner. Based on this data, we tried to identify the key symptoms or relevant thresholds for MCI diagnosis. The system presents the extracted knowledge in an easily understandable even for the users with minimal knowledge of statistics or data analysis.

Analýza a spracovanie textu

Analýza článků z českých zpravodajských serverů

Markéta Filipová, Jaroslav Kuchař

Fakulta informačních technologií
České vysoké učení technické v Praze
Thákurova 9, 160 00 Praha 6, Česká republika

{filipma4,jaroslav.kuchar}@fit.cvut.cz

Abstrakt. V dnešní době, kdy množství informací na internetu stále narůstá, se automatické zpracování a třídění dat stalo velmi oblíbeným oborem informačních technologií. Jednou z oblastí je i internetové zpravodajství. Cílem tohoto projektu je nástroj pokrývající celý proces pro základní analýzu článků z českých zpravodajských serverů. Projekt je zaměřen především na extrakci relevantních dat a jejich analýzu. V první části zahrnuje ale i související crawler, díky kterému je možné stáhnout články k analýze ze zpravodajských webů. V druhé části je ze stažených HTML stránek automaticky extrahován relevantní obsah článků a jejich další atributy. Třetí částí je pak textová analýza využívající existující postupy a nástroje, která se zaměřuje na extrakci pojmenovaných entit a analýzu sentimentu českého textu. Nad výslednými strukturovanými daty se lze dotazovat z různých pohledů a provádět tedy různé druhy experimentů.

Typ příspěvku: Aplikační příspěvek

Klíčová slova: zpravodajské servery, text mining, pojmenované entity, sentiment

1 Úvod

V českém prostředí existuje mnoho serverů zabývajících se zpravodajstvím a každý z nich se ve svém obsahu mírně liší. Informace jimi generované může být těžké analyzovat z několika důvodů. Články jsou umístěny na různých webových serverech a vyhledávat v nich je možné pouze pomocí dostupného vyhledávače. Důležité části článku, hlavně jeho obsah, je pak obklopen dalšími nežádoucími prvky, jako je šablona webu nebo reklamy. Není tedy jednoduché získat pouze relevantní informace. Kromě toho je článek psán v přirozeném jazyce a většinou k němu nejsou k dispozici žádné strukturované informace.

Cílem tohoto projektu je vytvořit takový nástroj [1], pomocí kterého by bylo možné získat z českých zpravodajských článků pouze relevantní informace a transformovat je do strukturované podoby, která umožní následnou analýzu zaměřenou na možnosti současných metod pro textovou analýzu českého textu.

2 Analýza zpravodajských článků

Proces získávání a analýzy dat je rozdělen do několika kroků. Předchází jim vlastní stažení jednotlivých článků ve formě HTML stránek z různých rubrik pro několik předních českých zpravodajských serverů (Novinky.cz, iDnes.cz, Aktuálně.cz a ParlamentníListy.cz), které je možné realizovat pomocí libovolného nástroje. Pro účely tohoto projektu jsme vytvořili crawler umožňující stažení článků z různých serverů dle zvolených rubrik a časových období.

2.1 Extrakce relevantních informací

Z HTML extrahujeme pouze nadpis článku, datum publikace, obsah a klíčová slova, což jsou prvky, které má k článku k dispozici většina českých zpravodajských serverů. Metoda extrakce využívá několika možností. V případě dostupnosti lze využít existující explicitní anotace (microdata). Pro ostatní případy navrhuje následující heuristiku extrakce relevantních dat.

Pro extrakci nadpisu využíváme faktu, že nadpis článku je vždy umístěn ve značce h1. Protože jich ale na stránce může být více, vybíráme nejlepší shodu dle Leveshteinovy vzdálenosti se značkou title, která nadpis článku obsahuje jako svou součást.

Datum publikace je ve stránce nalezeno pomocí regulárních výrazů a je přihlíženo k pozici nalezeného data na stránce vzhledem k pozici nadpisu a obsahu. Časové informace příliš vzdálené od hlavních sekcí jsou zahozeny.

Algoritmus pro detekci vlastního obsahu staví na principech existujících nástrojů [6] a algoritmů [2], které přizpůsobuje do oblasti zpravodajských serverů [1] a částečně českého prostředí s možností zobecnění. Pracuje tak, že v dokumentovém objektovém modelu stránky vyhledá textové uzly a odstraní ty, pro které odhadne na základě experimentálně ověřených parametrů, že obsahují nerelevantní obsah. V prvním kroku jsou odstraněny všechny prvky nacházející se nad nadpisem článku. Poté jsou z DOM vybrány všechny textové uzly, pro které je následně vypočítána hustota odkazů. Uzly nad stanovené procento (např. 70% textu je v odkazech) jsou označeny ke smazání. Následuje iterace přes všechny vybrané uzly, kde jsou označeny ke smazání další uzly podle toho, zda jsou sousední textové bloky označeny ke smazání či nikoli. Pokud je stanovený počet za sebou jdoucích uzlů označen ke smazání, je tak detekován konec stránky a zbylé uzly jsou označeny ke smazání. Poté se z DOM označené uzly odstraní. Celá stránka je následně pročištěna od uzlů, které neobsahují žádný text a od dalších nerelevantních prvků, jako např. tlačítka na sociální sítě.

Klíčová slova se v HTML většinou nacházejí ve specifických značkách (např. seznamy ul) se specifickými obsahy atributů (class, id...), jako jsou „tag“ či „keywords“.

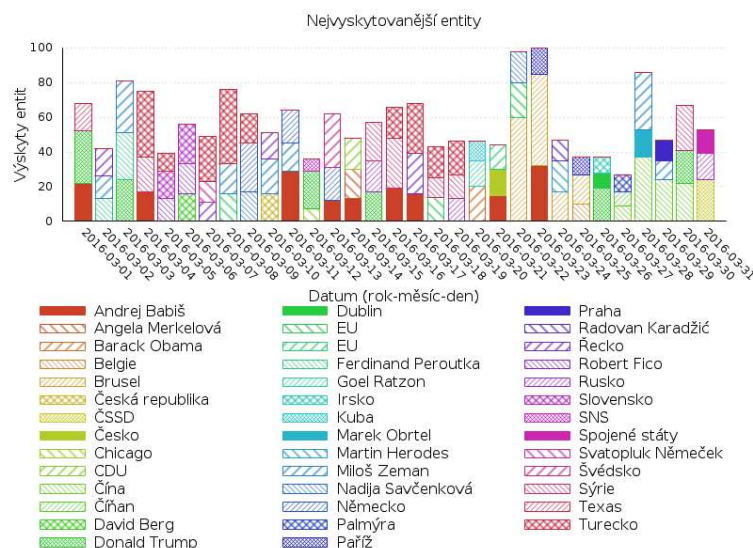
2.2 Extrakce pojmenovaných entit

Extrakce entit je založena na existujících programech NameTag [4] a MorphoDiTa [3]. Extrahovaný obsah článku používáme jako vstup do NameTagu, který v textu detekuje pojmenované entity a vrátí seznam entit ve tvaru, v jakém byly nalezeny v textu. Tyto

entity jsou pak převedeny do základního tvaru s využitím jednoduchého principu opakovaných výskytů v textech. Pokud je jeden z výskytů již v základním tvaru, využije se i pro ostatní. Výskyty každé nalezené entity jsou dále seskupeny. Pokud se v entitách vyskytuje osoba označená jménem a příjmením a dále pouze příjmením, jsou tyto různé tvary rovněž seskupeny do jedné entity. Pro každou entitu je pak zjištěno pomocí SPARQL dotazu, zda existuje její reprezentace na cs.dbpedia.org a pokud ano, je k entitě přiřazena její URI reference.

2.3 Analýza sentimentu

Pro analýzu sentimentu obsahu článku byl použit opět program MorphoDiTa [3] a navíc seznam českých emočně zabarvených slov SubLex [5]. Sentiment je v textu zjištěn jednoduchou slovníkovou metodou, kdy jsou slova textu článku převedena na svá lemmata a ta jsou následně porovnávána s emočně zabarvenými slovy. Každé nalezené pozitivní slovo má přiřazenu hodnotu 1, každé negativní pak -1, výsledný sentiment je součet nalezených emočně zabarvených slov. Určuje se jednak sentiment celého textu a jednak každé jeho věty.



Obr. 1. Nejvyskytovanější entity, březen 2016, iDnes.cz

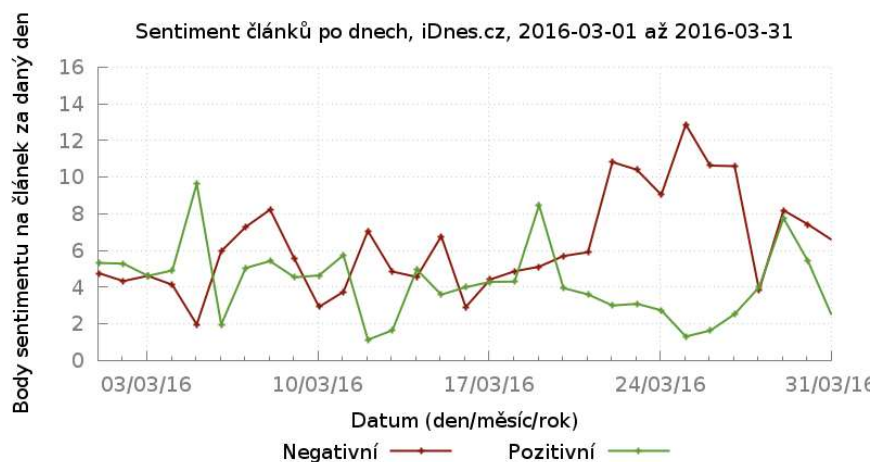
3 Evaluace

Dílčí části celého procesu jsme evaluovali na celkem šestnácti českých zpravodajských serverech a ručně anotovaných 44 článcích. Nadpis, datum a klíčová slova jsou extrahovány s plnou úspěšností. Samotný obsah je extrahován s průměrnou precision 0,993

a recall 1,0. Extrakce entit byla ověřena na vybraných 20 článcích s ruční anotací. Z celkových 772 výskytů bylo špatně přiřazeno či rozpoznáno 23 z nich (přibližně 3%). Analýzu sentimentu jsme ověřili na 32 ručně anotovaných článcích, kde došlo k neshodě v 7 případech (přibližně 21,9%). Výsledky jsou velmi ovlivněny svým malým rozsahem, samotnou kvalitou nástrojů NameTag, MorphoDita, SubLex a v případě analýzy sentimentu pohledem anotátora.

4 Experimenty

Experimenty jsme provedli na čtyřech českých zpravodajských serverech za období leden až duben 2016 (přes 17 tisíc článků). Na Obrázku 1 a Obrázku 2 se nachází příklad výstupu. Pro každý den jsou v grafu (Obrázek 1) entit zobrazeny tři nejvyskytovanější entit ze všech článků v daném dni za období březen 2016 ze serveru iDnes.cz. V grafu je vidět, že se v celém měsíci nacházejí jednak entity, které se poměrně často opakují (Turecko, Miloš Zeman, Evropská unie, Andrej Babiš...), a jednak entity, které se v měsíci příliš nevyskytují. Pomocí těch je možné detekovat zajímavé události. Z grafu je např. možné vyčíst, že 5. a 6. 3. 2016 jsou často vyskytované entity Slovensko a Robert Fico, což je z toho důvodu, že 5. 3. se na Slovensku konaly volby. Dále 22. a 23. 3. 2016 je znát velmi častý výskyt entity Brusel, což souvisí s teroristickými útoky v Bruselu v první zmíněný den. Rovněž v období 29. až 30. 3. je patrný vyšší výskyt entity Čína, což koresponduje s návštěvou čínského prezidenta v ČR v tomto období. Na grafu sentimentu (Obrázek 2) je zobrazen průměrný počet bodů sentimentu na jeden článek ze všech článků, které v daný den vyšly. Nejvýraznější je silná negativní vlna spojená s teroristickým útokem v Bruselu počínaje 22.3.2016.



Obr. 2. Sentiment článků (průměr bodů na článek) po dnech, březen 2016, iDnes.cz

5 Závěr

V projektu jsme vytvořili nástroje umožňující získání článků z různých českých zpravodajských serverů, extrakci důležitého obsahu, jeho základní textovou analýzu. Výstupy jsou uloženy ve strukturované podobě, která umožňuje provádět různé druhy pohledů a výstupů. Hlavním přínosem je upravený algoritmus extrakce relevantních informací a aplikace existujících metod pro analýzu získaných dat. Práce byla experimentálně ověřena na čtyřech předních zpravodajských serverech a dílčí nástroje byly evaluovány na datech poskytnutých anotátory.

Literatura

1. Filipová, M.: Nástroj pro analýzu článků z českých zpravodajských serverů. Diplomová práce, Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2016.
2. Kohlschütter, C., Fankhauser, P., Nejd, W.: Boilerplate detection using shallow text features. In Proceedings of the third ACM international conference on Web search and data mining, ACM, 2010, s. 441–450.
3. Straka, M. and Straková, J.: MorphoDiTa: Morphological Dictionary and Tagger, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, 2014, <http://hdl.handle.net/11858/00-097C-0000-0023-43CD-0>.
4. Straka, M. and Straková, J.: NameTag, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, 2014, <http://hdl.handle.net/11858/00-097C-0000-0023-43CE-E>.
5. Veselovská, K. and Bojar, O.: Czech SubLex 1.0, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague, 2013, <http://hdl.handle.net/11858/00-097C-0000-0022-FF60-B>.
6. Readability GitHub.com [online]. 2016, [cit. 2016-04-26]. Dostupné z: <https://github.com/luin/readability>

Annotation:

Analysis of Czech news articles

Nowadays, when the amount of information on the internet continues to grow, automatic processing and analysis of data has become a very. Online news service is one of the domains in which a significant amount of diverse as well as similar information exists. The goal of this work is to create a tool for analysis of Czech news articles. The first part is a crawler which allows downloading articles for analysis from news servers. In the second part, relevant content of articles and their other attributes are extracted from downloaded HTML pages. The third part is a text analysis for which modules for extraction of named entities and for sentiment analysis of Czech texts have been created. We performed experiments on four of the most visited Czech news portals. The results show that the presented approach is suitable for analysis of news articles.

Interactive Evolution and Poem Models in Haiku Poetry Generation

Miroslava Hrešková, Kristína Machová

Faculty of Electrical Engineering and Informatics
Technical university of Košice
Letná 9, 042 00 Košice, Slovakia

miroslava.hreskova@student.tuke.sk,
kristina.machova@tuke.sk

Abstract. Innovative field of computational creativity is focusing on developing algorithms that are capable of creating outputs that would be considered creative. Language provides a lot of opportunities for creativity, so the article describes and compares two different approaches to generating haiku poetry. First approach proposes using evolutionary algorithm and human as a fitness function in the evolution. Second described approach composes poems based on haiku models that were extracted from haiku database. The goal is to create poems, considered by humans as understandable and with aesthetic value.

Contribution type: Work-in-progress paper

Keywords: computational creativity, haiku poetry, natural language generation, poetry generation

1 Introduction

Computational creativity is dedicated to studying and creating systems that can be considered creative. Its goal is to study human creativity and make systems that are capable of creating outputs that would be considered creative, if the same output were produced by human. This article proposes and compares two approaches for computational creativity algorithms. Both approaches aim to create haiku poems.

Haiku is a genre of poetry that has its origin in Japan, but was also adapted to other languages as well. This article describes generation of haikus in English language. Traditional haiku poem consists of three lines with fixed syllable count for each verse (5-7-5 syllable pattern). Regarding content, main theme of haiku poem is nature and it aims to capture a feeling.

First approach leverages interactive evolutionary computation, artificial intelligence method combining evolutionary algorithms with human as a fitness function to achieve personalised outputs created and modified during evolution.

Second approach aims to implement certain level of knowledge about haiku poems, mostly its structure and content. It relies on creating models of poems from large haiku poems created by human authors.

Analysis and description of several related works can be found in section 2. Next sections are dedicated to description of the proposed approaches for haiku poetry generation, as well as listing of several example outputs. Last part of the article compares both proposed approaches and describes further work.

2 Related works

Some of existing applications of computational creativity with focus on generating poetry are described in this section.

2.1 Poetry Generation

System POEVOLVE [3] generates poems using evolutionary computation with evaluation function implemented as a neural network. This neural network was trained on data obtained from human evaluations of poems and it is used to solve problem with human fatigue caused by evaluating lots of individuals. This approach also helps with speeding up the poem creation process. Using interactive evolution in haiku generation was an inspiration for first described approach in this article.

System called WASP [4] generates different genres of poetry. From input data, which consist of a set of words and a set of reference verse patterns, it creates poems by using the words sent to the system as input. The output satisfies constraints of input reference patterns.

Another interesting poetry generator is system called Tra-La-La Lyrics [5]. Input to the system is a melody. The system creates poem as a lyrics for the input melody. The poem generation is based on observation that strong beats in the song are associated with the lexical stress in the words. It is very interesting application for creation of song lyrics.

2.2 Haiku Generation

Many haiku generators can be found on the internet. Poetry engine¹ is an example of such freely available generator. Its outputs are just random selection of words, such poems rarely have meaning and are made to entertain users.

Generator² is another example of poetry engine available through web. To generate poem, it uses pre-defined models of sentences to create haiku poems. Words are randomly chosen from thematic dictionary based on its part of speech. Using models and

¹ Random Haiku Generator. [online, cited 10.09.2016]. Available at <http://www.randomhaiku.com/>

² Peter's Haiku Generator. [online, cited 10.05.2016]. Available at <http://peterhoward.org/haikugen/frameset1.htm>.

having a database of words commonly used in haiku poems inspired second described approach to haiku generation in this article.

3 Haiku Poetry Generation with Interactive Evolution

Evolutionary algorithm is used to create new poems and leverages human evaluation as fitness function. The proposed approach [1] was implemented as web application.

Generation of haiku poems with interactive evolutionary computation was chosen because the user's subjective preference is the most important task when generating natural language, especially poetry.

10 users participated in experiment that was carried out to evaluate the performance of the application. 50% of the participants were satisfied with poems in final generation.

3.1 Haiku Evolution

The source for creating initial population is haiku corpus. It is a database consisting of haiku poems created by human authors. AhaPoetry³ and DailyHaiku⁴ were used as the source for the corpus.

Verse is considered the basic element of the poem, so each individual is structured as a collection of 3 verses. Each population contains 10 individuals. Population size was chosen experimentally – several experiments with real users were carried out and their opinion was taken into account when choosing the size of population for evolution.

The number of evaluations that IEC can receive from one human user is limited by user fatigue. This is a big disadvantage of interactive evolution, because smaller search space is explored.

At first, every poem in population has the same fitness value which is modified based on human evaluation. Then, haiku poems with positive and neutral feedback from human evaluator are chosen to reproduce. Cross-over with two and three parents are genetic operators designed to create new individuals in reproduction stage of evolutionary cycle.

3.2 Example Outputs

Several poems generated by system:

*picking wildflowers
the early spring sun
in my hand*

*cherry blossoms
the ant carries only*

³ Aha Poetry. [online, cited 10.05.2016]. Available at <http://www.ahapoetry.com/aadoh/h_dictionary.htm>.

⁴ Daily Haiku. [online, cited 10.05.2016]. Available at <<http://www.dailyhaiku.org/>>.

its papery leaves

*a winter bracken
in full bloom he photographs
his fat wife*

4 Haiku Poetry Generation with Poem Models

Haiku generation with poem models constructs poems from words based on their part of speech and number of syllables. It was designed to avoid the problems in interactive evolution approach. The approach was implemented as web application.

No experiments have been conducted to evaluate performance of the application as of yet. The outputs were compared to outputs generated by the interactive evolution. The comparison can be found in section 5.

This approach [2] ensures that final poem will conform to required syllable pattern by using syllable counting algorithm during poem creation. With using dictionary extracted from haiku poem, it aims to create poems with haiku-specific words and by this to take into account also content criteria.

4.1 Haiku Generation

Word is considered the basic element of the poem and poem is constructed from words based on word metadata. Pattern for word selection (to fill the model with words) is defined by poem model.

In data preparation phase, dictionary is created. For creating dictionary, haiku corpus (consisting of the same haiku poems as haiku corpus used in interactive evolution of poems) is used.

Dictionary consist of all words from poems in haiku corpus. Each word in dictionary is defined by 3 properties:

- word itself
- part of speech (metadata)
- syllable count (metadata)

Haiku corpus is also used for poem model extraction. Poem model consists of list of parts of speeches and list of syllable counts. Poem model is constructed from every poem in haiku corpus. Unique haiku models with frequency (occurrence number) 5 are kept as haiku specific. Frequency constant was chosen experimentally.

Words from dictionary are selected into the poem based on part of speech and number of syllables. The poem is displayed for evaluation by human user to determine the performance of the system.

4.2 Example Outputs

Several poems generated by system:

*always helpless
nearby responsibly shrivels
a greedy banyan*

*in flawless garden
mysterious dahlia
quietly complains*

*something which blossoms
conspire according cow
dry shadow outside*

5 Comparison of Proposed Approaches

Noticable problem of poems generated with interactive evolution is that not all of the poems conform to 5-7-5 syllable pattern. The reason is that not every poem in haiku corpus follows the syllable pattern and it reflects into generated poems as well. To avoid issue with not conforming to formal haiku criteria, approach for generating haikus using poem models was proposed. To make sure that syllable count rule is followed, it takes into account syllable number in process of poem creation.

Both approaches create haiku poems that contain vocabulary related to nature or vocabulary used to express emotion. The haiku content is provided by using large corpus with haiku poems written by human authors.

Sometimes, both systems create less meaningful poems. In case of interactive evolution, this happens when two poems with different topics and/or with different emotion are selected to cross-over. As for poem models, the system does not have any further knowledge on how to select words from dictionary into poem, so that it would choose words related by sentiment and topic.

6 Conclusion

When comparing generated poems of both systems, generating haiku poetry with poem models creates better poems from the formal point of view and also in terms of poem content. The reason is that it is easier to combine words than whole verses

The application creating poems by using poem models will be later made accessible to wide public in order to evaluate and test the performance of poetry generator with real users and continuously improve it.

References

1. HRONCOVÁ, Miroslava; MACHOVÁ, Kristína. Generating Haiku Poems Using Methods of Artificial Intelligence. In: WIKT 2015. Košice TU, 2015. p. 96-102
2. HRONCOVÁ, Miroslava; MACHOVÁ, Kristína. Haiku Poetry Generation. In: 16th Scientific Conference of Young Researchers: Proceedings from Conference. Faculty of Electrical Engineering and Informatics Technical University of Košice, 2016
3. LEVY, Robert P. A computational model of poetic creativity with neural network as measure of adaptive fitness. In: Proceedings of the ICCBR-01 Workshop on Creative Systems. 2001.
4. GERVAS, Pablo. Wasp: Evaluation of different strategies for the automatic generation of spanish verse. In: Proceedings of the AISB-00 symposium on creative & cultural aspects of AI. 2000.
5. OLIVEIRA, Hugo Gonçalo. CARDOSO, F. Amilcar. PEREIRA, Francisco C. Exploring different strategies for the automatic generation of song lyrics with tra-la-lyrics. In: Proceedings of 13th Portuguese Conference on Artificial Intelligence, EPIA, pp. 57-68. 2007.

Slovenský stemmer emocionálnych slov

Zuzana Nemčíštinová, Martin Mikula, Kristína Machová

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

`zuzana.nemcisinova@student.tuke.sk, {martin.mikula,
kristina.machova}@tuke.sk`

Abstrakt. Emocionálne slová hrajú významnú úlohu v procese klasifikácie názorov a analýzy sentimentu. Ich nájdenie a identifikácia v texte je preto veľmi dôležitá. Po fáze vyhľadávania nasleduje fáza ich spracovania. Pre zjednodušenie spracovania a zlepšenie výsledkov je vhodné použiť určitú formu úpravy týchto slov, či už pomocou stemmovania alebo lematizácie. Táto práca je venovaná vytvoreniu slovenského stemmera pre emocionálne slová, ktoré sú následne použité pre klasifikáciu emócií. Stemmer je založený na gramatických pravidlách slovenského jazyka a dokáže na základe zadaných prípon, predpon a pravidiel previesť slovo v akomkoľvek morfológickom tvare na jeho kmeňový tvar bez toho, aby vedel, o aký slovný druh ide. Medzi slovné druhy, ktoré je stemmer schopný spracovať patria podstatné mená, prídavné mená a slovesá nesúce akúkoľvek emóciu. Na testovanie sme použili slovník obsahujúci 17 872 vyskloňovaných slov. Stemmer dosiahol celkovú presnosť 98,1%.

Typ príspevku: Výskumný príspevok

Kľúčové slová: stemmer, predpona, prípona, slovenský jazyk, kmeň slova

1 Úvod

Každý z nás prichádza každý deň do kontaktu s internetom, ktorý používame na komunikáciu, zábavu, prácu, či vyhľadávanie informácií. V dnešnej dobe sa na internete nachádza veľké množstvo dát, ktoré je zložité manuálne prehľadávať. Z tohto dôvodu sa dostávajú do popredia mechanizmy slúžiace na získavanie a vyhľadávanie informácií. Práve v tejto oblasti sa používajú stemmovacie algoritmy, ktoré majú najväčšie uplatnenie najmä vo webových vyhľadávačoch, ktorým je napríklad Google. Okrem webových nástrojov je možné tieto algoritmy využiť napríklad pri spracovaní prirodzeného jazyka. A práve pri počítačovom spracovaní je slovenský jazyk veľmi zložitý. Slovenčina patrí k jazykom s bohatou morfológiou, čo znamená, že slovo v texte môže mať rôzne tvary, a toto je hlavný dôvod, prečo slovenský jazyk v porovnaní s inými jazykmi značne zaostáva. Veľké množstvo výnimiek a tvarov slov, vznikajúcich pri skloňovaní, výrazne komplikuje prácu so slovenským textom. Z toho dôvodu je veľmi náročné vytvoriť algoritmus, ktorý prevedie jednotlivé slová na ich základný tvar.

2 Problematika stemmovania v slovenskom jazyku

Existuje niekoľko druhov stemmovacích, respektíve lematizačných algoritmov, ktoré pracujú na rôznych princípoch a dosahujú rôzne výsledky. Tieto algoritmy majú rôzne výhody a nevýhody. Niektoré dosahujú dobré a kvalitné výsledky, ale sú časovo a výkonnostne náročné. Ostatné dosahujú menej kvalitné výsledky, ale ich výhodou je práve rýchlosť spracovania výsledkov [1].

Slovenský stemmer podstatných mien bol implementovaný do vyhľadávacieho systému Lucene. Stemmer bol inšpirovaný pravidlami pre ruský stemmer. Ide o program, ktorý previedol vyskloňované podstatné mená na ich základný tvar, na základe odstraňovania prípon. Táto aplikácia dosahovala úspešnosť približne 90 %. Stemmer bol testovaný na dvoch článok a výstupom boli dvojice, a to slovo a jeho koreň. [2]

Slovenský stemmer slovenských priezvisk a názvov ulíc bol taktiež implementovaný do Lucene. Tento stemmer dokáže odstrániť prípony, ktoré obsahujú slovenské priezviská a názvy ulíc. Aplikácia bola testovaná na 70 slovách obsahujúcich názvy ulíc a priezviskách v rôznych tvaroch. Tento stemmer dosahoval takmer 99 % úspešnosť, pričom nesprávne vyhodnotil iba jedno meno cudzieho pôvodu. [3]

Tvaroslovník je program, ktorý bol vytváraný na UPJŠ v Košiciach. Ide o databázu, ktorá obsahuje 30 000 000 tvarov slovenských slov. Tieto slová sú uložené v databáze formou textových súborov. Každé slovo obsahuje záznam o tom, o aký slovný druh ide a aké sú jeho gramatické kategórie. Tieto súbory boli uložené do databázy, čím bolo možné použiť tento program na lematizáciu alebo na získanie všetkých tvarov slov pre dané slovo. Program používa na stemmovanie predlohu, čiže na základe porovnávania slova a predlohy hľadá základný tvar pre dané slovo. Tak ako každý z lematizátorov slovenského jazyka, ani tento nie je bezchybný. Keďže textové súbory, ktoré boli vložené do databázy, vytváralo niekoľko študentov a nie všetci pristupovali k svojej práci zodpovedne, z tohto dôvodu sa v tejto práci vyskytlo aj množstvo chýb. Rýchlosť lematizácie dosahuje v priemere 132 slov/sekunda. Táto databáza všetkých tvarov slov sa často využíva aj v mnohých ďalších projektoch. [4] [5]

3 Stemmer pre klasifikáciu emócií

Slovenský stemmer pre klasifikáciu emócií používa na stemmovanie emocionálnych slov algoritmus odstraňovania prípon a predpôn. Keďže slovenčina patrí k jazykom s bohatou morfológiou, často v nej dochádza ku tvorbe výnimiek. Nakoľko väčšina emocionálnych slov má pravidelné stupňovanie, bol pre nás problém ohľadom výnimiek zanedbateľný. Medzi výhody tohto prístupu patrí vysoká presnosť, pri aplikovaní správnych pravidiel a rýchlosť. Tento stemmer bude následne implementovaný do algoritmu na analýzu sentimentu, kde by mal výrazne urýchliť dobu klasifikácie oproti stemmovaniu založenému na slovníku.

Pri tvorbe prípon sme postupovali spôsobom, pri ktorom sme najskôr získali všetky prípony vznikajúce pri skloňovaní prídavných mien, podstatných mien a pri časovaní sloves. Tieto prípony sme získali z Pravidiel slovenského pravopisu. Prípony sme spojili a odstránili duplicitné. Nakoniec sme získali pole koncoviek, ktoré obsahuje 130

prípon (Tab 1). Z predpôn sme vybrali najmä predpony cudzieho pôvodu. Niektoré z nich slúžia pri tvorbe slov skladaním, napríklad dobro-srdečný. Tab. 1 znázorňuje aj 13 predpôn, ktoré v prípade, ak sa nachádzajú v slove, tak budú pri stemmovaní odstránené.

Tab 2. Tabuľka obsahujúca všetky použité predpôn a prípon.

| | |
|----------|---|
| PREDPONY | <i>dis-, kilo-, homo-, mega-, malo-, polo-, seba-, poly-, hypo-, ultra-, infra-, dobro-, hyper-</i> |
| PRÍPONY | <i>encoch, -encami, -ujete, -ujeme, -ovalo, -ovali, -ovala, -eniec, -encom, -atami, -atach, -ujuc, -ujte, -ujme, -ujes, -ujem, -ovia, -ovat, -oval, -iete, -iemu, -ieme, -ieho, -iami, -ialo, -iali, -iala, -iach, -ence, -ejuc, -ejte -ejme, -atom, -atam, -ajuc, -ajte, -ajme, -ymi, -ych, -ulo, -uli, -ula, -uju, -uje, -ovi, -och, -ite, -iou, -iom, -imi, -ime, -ilo, -ili, -ila, -ich, -iev, -iet, -ies, -ien, -iem, -iel, -iej, -iat, -iam, -iac, -ete, -emu, -eme, -elo, -eli, -ela, -eju, -eho, -aty, -atu, -ati, -ate, -ata, -ami, -ame, -alo, -ali, -ala, -aju, -ach, -ym, -ut, -us, -ul, -uj, -uc, -te, -ov, -ou, -om, -ol, -ok, -mu, -mi, -me, -lo, -li, -la, -iu, -it, -is, -io, -im, -il, -ii, -ie, -ia, -ho, -es, -en, -em, -el, -ej, -at, -as, -am, -al, -aj, -ac, -y, -u, -o, -i, -e, -a.</i> |

Algoritmus na stemmovanie emocionálnych slov, sa skladá z nasledujúcich častí:

Vloženie a predspracovanie textového súboru, respektíve slova:

- Vloženie a načítanie slova alebo textového súboru.
- Konverzia veľkých písmen na malé a odstránenie diakritiky.
- Algoritmus ošetruje dĺžku slova a v prípade, ak sa slovo skladá z 3 písmen a súčasne končí na niektorú zo spoluhlások: „-j, -l, -m, -s, -t, -v, -z“, bude o tom používateľ oboznámený a algoritmus vypíše na výstup výsledné slovo v pôvodnom stave inak pokračuje ďalej.

Odstránenie predpôn:

- Ak slovo začína na predponu „dis-“, tak táto predpona bude odstránená.
- Ak je dĺžka slova väčšia ako 4 písmena a začína na niektorú z predpôn: „homo-, kilo-, malo-, mega-, polo-, seba-, poly-, hypo-“, tak prípona bude odstránená.
- Ak je dĺžka slova väčšia ako 5 písmen a začína na predponu: „dobro-, infra-, ultra-, hyper-“, tak bude táto predpona odstránená. Dĺžka pri predponách sa ošetruje z toho dôvodu, že predpona môže byť odstránená aj vtedy, keď znázorňuje samotné slovo. Napríklad predpona dobro-, môže znázorňovať aj samotné podstatné meno dobro v nominatíve a v prípade, ak by nebola ošetrená dĺžka slova, toto slovo by bolo odstránené.

Odstránenie prípon:

- Ošetrovanie dĺžky slova a nastavenie dĺžky prehľadávanej prípony.

- V prípade, ak je dĺžka slova väčšia ako 7 písmen, algoritmus nastaví začiatok prehľadávanej prípony ako rozdiel dĺžky slova a najdlhšej prípony. To znamená, $k = \text{dĺžka slova} - 6$. Napríklad, ak sa slovo skladá z 8 písmen, tak k vypočítame ako $8 - 6 = 2$, a v takomto prípade začne slovo prehľadávať od 3. písmena.
- V prípade, ak je slovo kratšie ako 7 písmen, premenná k bude nastavená na 2. Je to z toho dôvodu, že slová kratšie ako 2 písmená nebudú vôbec stemmované.
- Začiatok cyklu *for*, ktorý slúži na prehľadávanie slova a porovnávanie prípon.
- Algoritmus prehľadáva slová od najdlhšej novej prípony a porovnáva ju s vopred zadefinovanými príponami (pole koncoviek).
- V prípade, ak bola nájdená zhodná prípona, algoritmus ju odstráni, inak pokračuje v prehľadávaní a porovnávaní prípon. Ak sa nenájde žiadna zhodná prípona, bude to znamenať, že slovo nemá príponu a predstavuje už hľadaný kmeňový tvar.

Odstránenie stupňovania:

- V tomto kroku sa ošetruje 3. stupeň prídavných mien, to znamená, že ak slovo po odstránení predchádzajúcej prípony končí na „-s“ a začína predponou „*naj-*“, algoritmus túto predponu odstráni a informuje o tom používateľa.
- V prípade, ak slovo končí na príponu „-*ejš*“, bude aj táto prípona, používaná v 2. a 3. stupni prídavných mien, odstránená.

4 Testovanie a vyhodnotenie aplikácie

Aplikácia slovenský stemmer pre klasifikáciu emócií bola testovaná prostredníctvom slovníka sentimentálnych slov. Slovník obsahuje 17 872 slov, z toho 2 473 podstatných mien, 11 293 prídavných mien a 4 106 sloves. Ostemované slová boli manuálne porovnávané s kmeňmi určenými expertom. Na vyhodnotenie presnosti sme použili vzorec na výpočet presnosti, konkrétne sme vychádzali z tohto vzorca: $p = \frac{TP}{(TP+FP)}$,

kde TP, predstavuje správne ostemované slová aplikáciou a FP znázorňuje nesprávne ostemované slová aplikáciou. Presnosti pre jednotlivé slovné druhy, ako aj celková presnosť sú popísané v tabuľke Tab. 2.

Tab. 2 Tabuľka obsahujúca presnosť stemovania pre jednotlivé slovné druhy.

| Slovný druh | ostemované slová | neostemované slová | všetky slová | presnosť (%) |
|----------------|------------------|--------------------|--------------|--------------|
| podstatné mená | 2390 | 83 | 2473 | 96,64 |
| prídavné mená | 11137 | 156 | 11293 | 98,61 |
| slovesá | 4011 | 95 | 4106 | 97,68 |

Táto vysoká presnosť je podľa nášho názoru spôsobená odstraňovaním diakritiky hneď na začiatku stemmovania. Keďže pri skloňovaní slov dochádza ku zmene diakritických

znamienok v poslednej slabike slova, tak v prípade, ak by sme sa zaoberali aj diakritikou, tento stemmer by dosahoval podstatne nižšiu presnosť.

Ďalším z dôvodov, prečo je táto presnosť taká vysoká, môže byť aj orientácia stemmera na emocionálne slová. Je možné, že v prípade implementácie tohto algoritmu na neutrálne slová by mohlo dôjsť k zníženiu presnosti, a to z dôvodu, že podstatné mená, v ktorých dochádza ku výnimkám v skloňovaní, predstavujú najmenšiu časť z testovacej množiny a nachádza sa tam zanedbateľný počet slov, ktoré tvoria výnimku (zlo → zle, neha → nieh).

5 Záver

Stemmer, vytvorený v tejto práci, predstavuje druh stemmeru zameraný na stemmovanie emocionálnych slov. Stemmer funguje na princípe odstraňovania prípon a predpôn, a to na základe pravidiel slovenského jazyka. Od ostatných stemmerov sa odlišuje najmä použitím predpôn, zahŕňa predpony, ktoré sa vyskytujú najmä v cudzích slovách vyjadrujúcich emóciu, prípadne pri iných emocionálnych slovách, napríklad aj pri slovách tvorených skladaním, príkladom je slovo *dobrosrdečný*. Aplikácia dosahla veľmi dobré výsledky, ktoré boli testované na 17 872 emocionálnych slovách. Po otestovaní sme zistili, že aplikácia dosahuje priemernú presnosť 97,64%. Ide o vysokú presnosť, ktorá môže byť spôsobená odstraňovaním diakritiky, orientáciou na emocionálne slová a testovacia množina s malým počtom slov, reprezentujúcich výnimky v slovenčine.

Pod'akovanie. Tento príspevok vznikol s podporou Vedeckej grantovej agentúry Ministerstva školstva, vedy, výskumu a športu Slovenskej republiky v rámci projektu č. 1/0493/16 „Metódy a modely pre analýzu prúdov dát“.

Literatúra

1. Šanda, P.: Určení základního tvaru slova. Diplomová práce, Brno, (2011). [cit. 2016-03-06]. Dostupné na internete: <https://dspace.vutbr.cz/bitstream/handle/11012/6088/DP_Ur%C4%8Den%C3%ADZ%C3%A1kladn%C3%ADhoTvaruSlova.pdf?sequence=1>.
2. Pífková, H.: Slovenský stemmer. [cit. 2016-02-20]. Dostupné na internete: <http://vi.ikt.ui.sav.sk/Projekty/Projekty_2008%2F%2F2009/Hana_Pifkova%C3%A1_-_Stemer>.
3. Balocký, S.: Slovenský Stemmer. [cit. 2016-02-20]. Dostupné na internete: <http://vi.ikt.ui.sav.sk/Projekty/Projekty_2008%2F%2F2009/Stislav_Balocky?highlight=stemmer>.
4. Horváth, T.: Informačné Technológie – Aplikácie a Teória. [cit. 2016-02-24]. Dostupné na internete: <<http://itat.ics.upjs.sk/proceedings/ITAT2012%20-%20Zbornik%20prispievkov%20-%20tlacova%20verzia.pdf>>.
5. Turlíková, L.: Tvaroslovník a jeho využitie vo vetnom rozboře. [cit. 2016-02-24]. Dostupné na internete: <<http://web.ics.upjs.sk/svoc2009/prace/8/Turlikova.pdf>>.

Annotation:

Slovak stemmer for emotional words

The emotional words have important role in opinion classification and sentiment analysis. It is very important to find and identify them. After identification of these emotional words, it is very important to process them correctly. We can use stemming or lemmatization to process them. This paper is focused on creation Slovak stemmer for emotional words, which can be used for opinion classification. Our stemmer is based on grammatical rules, it can remove prefixes and suffixes and it can find stem of the word. We can process mainly adjectives, nouns and verbs which contain emotions. The stemmer was tested on 17 872 words and achieved accuracy 98,1%.

Text Analyzing of Aviation Safety Reports

Lama Saeeda, Petr Křemen, Marek Štumper

Czech Technical University in Prague, Czech Republic

{saeeda.lama@fel, petr.kremen@fel, stumpmar@fd}.cvut.cz

Abstract. Aviation safety reports start to play a crucial role in understanding incidents and accidents in the aviation safety field. Automated text processing is necessary for simplification of the safety reporting process. This task can be achieved in different ways, such as statistical, non-statistical or a combination of these techniques. In this paper, we are mainly focusing on non-statistical ones, by introducing our text processing scenario. We start with indexing of various aviation safety vocabularies that we are using as a backbone for this task. Next, the golden standard corpus is prepared, including the testing process of several Linked Data Knowledge Extraction tools, with respect to a domain-specific vocabulary. Then, choosing the most accurate entity annotation tools and making them work together, as well as with other features that we added, taking into consideration some very specific terms and abbreviations used in aviation field. The ultimate goal is to build a tool that will integrate several techniques inside in order to provide high precision reports' annotations in aviation safety domain.

Contribution type: Research paper

Keywords: text analyzing, ontology, aviation safety

1 Introduction

Initial incident and accident reports are the best sources of information for extracting the most important knowledge to feed the preliminary¹ reports' building process. Initial reports are usually a free-form text, describing the incident or the accident, along with a small set of metadata (mostly concerned with the time, the location and the equipment involved [1]). The automatic analyzing process of such reports is challenging, because they are usually short, and they contain a lot of aviation-specific terms and abbreviations. Recently, some entity recognition tools have appeared. As mentioned in [2], if a custom vocabulary can be loaded (configured or programmed) into the tool, it significantly improves the recognition of the entities. For that, the selection of the tools was focused only on the ones that are combining both Natural Language Processing (NLP)

¹ Preliminary report is created by safety department of an organization and sent to the authority.

and Linked Data capabilities, and allow building custom indexes depending on the domain-specific vocabularies.

This paper describes our experience gathered during analysis of aviation safety reports in the Czech environment, as well as evaluation of our pipeline on selected reports.

2 Description of the corpus

In order to evaluate the pipeline, we had to create gold standard corpus. It mainly consists of initial safety reports in the aviation-safety domain. Experts in aviation domain manually annotated domain terms (entities) in each report with respect to huge controlled domain-specific vocabularies. Technically, they used the General Architecture for Text Engineering (GATE) tool². We need this kind of corpus for the evaluation process of the tool, as well as augmenting our aviation ontology³ with more terms and relations in our future work.

3 Aviation Safety Text Analyzing Tool

The text processing helps in building the preliminary safety report based on the initial ones. The initial reports usually contain very basic information (written in a natural language) about the specific accident or incident, such as the safety occurrence participants, place, time, etc..

For better text understanding and entity recognizing, many techniques were introduced. Some are ontological-based, where the ontologies and the other knowledge resources are widely used to aid the recognition in special domain texts, besides by the linkage from the text back to the ontologies, we can achieve better understanding and gain additional knowledge. Also the statistical-based entity recognition models with its various algorithms can overcome some of the shortcomings of the other techniques.

In order to detect entities in such reports, several entity recognition tools were tested. We described a portfolio of objects and events or artifacts that are important for the safety reporting. We are showing the roadmap how the task of detecting the most important information in text works and make use of it in the aviation safety reporting tool. These tools are mentioned in the next paragraphs.

3.1 Apache Stanbol

Apache Stanbol provides the ability to work with custom vocabularies and creating custom indexes upon it, which is necessary for being able to detect various types of entities, and to detect and work with concepts from a specific domain [3]. It also comes with a list of enhancement engines implementations, with the ability to build a specific

² <https://gate.ac.uk/>

³ <https://www.inbas.cz/aviation-safety-ontology>

one to get the most benefit out of the tool [4]. This allowed us to build a chain of enhancement engines that fits perfectly to the aviation-safety concepts detection.

3.2 DBpedia Spotlight

DBpedia Spotlight offers three basic functions, Annotate, Disambiguate and best K-Candidates. It can be accessed from a REST Web Service and from a user interface on the Web [5]. It also offers creating a Spotlight model on the user's own server through an internationalization process, to model occurrences of resources with the context in which they have been mentioned.

Indexing process and building a customized index according to the aviation ontology is an intensive task with DBpedia Spotlight. It needs extra efforts to extract surface forms and valid URIs from the gold standard corpus and then, build the dictionary-based spotter from them [6].

3.3 Customized techniques

Some of the artifacts that we defined are hard to be detected by the previously mentioned tools, in spite of their indexing capabilities and their ability to detect mentions from the specific terminology. For these specific terms, we are using different detection techniques using the advantage of pre-knowledge of its rules. For detecting aircraft call signs, for instance, regular expressions are used, taking into consideration the rules of different possible formats for call sign representation [7].

The output of Apache Stanbol, DBpedia Spotlight and the customized techniques were parsed, merged and optimized in a RESTful web service. The service outputs the entities that are detected, with their proper mapping to the aviation ontology.

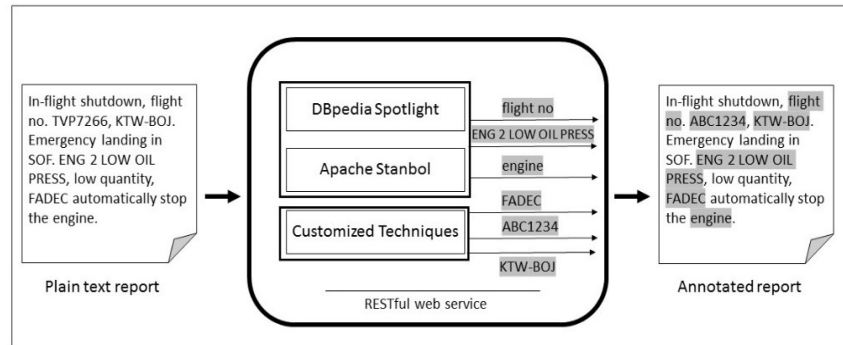


Fig. 1. The processing pipeline

4 The evaluation process

A testing framework was developed to achieve the evaluation task. The main process is to parse the output of the tool and compare it to the expected annotations in the corresponding document in the gold standard.

The statistics that were observed were basically the True Positives (TP), where the cases were positive and predicted positive, the False Positives (FP), where the cases were negative but predicted positive, and the False Negatives (FN), where the cases were positive but predicted negative. These statistics are then used to calculate the usual precision, recall and the F1 measures.

Table 1. Evaluation results for samples of reports

| | Precision | | Recall | | F1 | |
|----------|------------------------|---------------------------|------------------------|---------------------------|------------------------|---------------------------|
| | With custom vocabulary | Without Custom vocabulary | With custom vocabulary | Without Custom vocabulary | With custom vocabulary | Without Custom vocabulary |
| Report1 | 0.5 | 0 | 0.095 | 0 | 0.160 | 0 |
| Report2 | 0.3125 | 0.071 | 0.555 | 0.111 | 0.4 | 0.091 |
| Report3 | 0.625 | 0.286 | 0.435 | 0.174 | 0.513 | 0.216 |
| Report4 | 1 | 1 | 0.278 | 0.167 | 0.435 | 0.286 |
| Report5 | 0.5 | 0 | 0.1 | 0 | 0.1667 | 0 |
| Report6 | 1 | 1 | 0.583 | 0.417 | 0.737 | 0.588 |
| Report7 | 0.5 | 0.5 | 0.182 | 0.182 | 0.267 | 0.267 |
| Report8 | 0.444 | 0.083 | 0.381 | 0.048 | 0.410 | 0.061 |
| Report9 | 0.667 | 0 | 0.286 | 0 | 0.399 | 0 |
| Report10 | 0.5 | 0.4 | 0.375 | 0.25 | 0.428 | 0.308 |
| Report11 | 0.667 | 0 | 0.182 | 0 | 0.286 | 0 |
| Report12 | 0.8 | 0.5 | 0.470 | 0.118 | 0.592 | 0.19 |
| Report13 | 1 | 1 | 0.148 | 0.148 | 0.258 | 0.26 |

As we can observe from the evaluation statistics of arbitrary samples of the corpus, the precision scores high rates in the most of the cases. It even reaches to 100% rate for some reports. On the other hand, the recall scores low rates. This affects the F1 measure to be lower for the most of the reports. However, in our case, F1 measure might not be the best evaluation criteria. As mentioned previously, our ultimate goal is to achieve high precision annotations for the aviation safety reports in order to be directly used in practice.

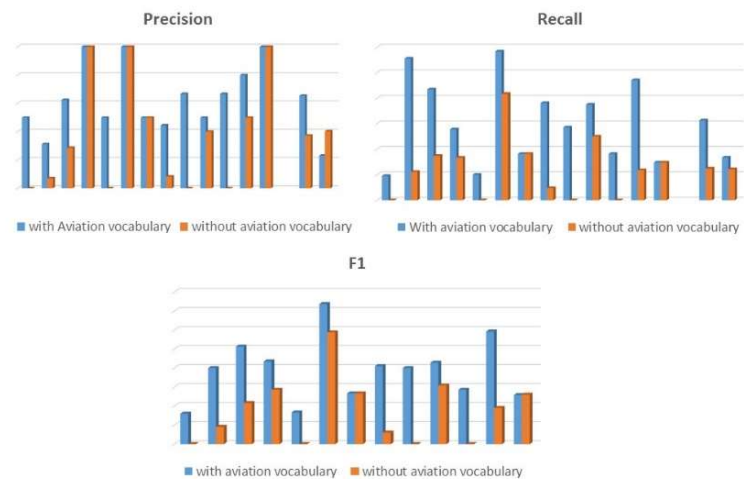


Fig. 2. Reports' evaluation charts

| | |
|---|--|
| Flight had a prolonged loss of communication over Swiss territory. Zurich radar informed at 09.28.39 about loss of contact and that also no contact on 121,5 MHz could be established. Geneva informed 09.46.35 that ABC1234 has contacted them. Length of loss of comm. is approx. 11 minutes. | Flight had a prolonged loss of communication over Swiss territory. Zurich radar informed at 09.28.39 about loss of contact and that also no contact on 121,5 Mhz could be established. Geneva informed 09.46.35 that ABC1234 has contacted them. Length of loss of comm. is approx 11 minutes. |
|---|--|

Fig. 3. Sample of manually annotated report (R11) by experts

Fig. 4. Sample of automatically annotated report (R11) by the tool

Table 2. Entities detected and their types according to the Aviation ontology⁴

| Entity Name | Entity Type |
|---------------------------------|-------------|
| Flight | Event |
| prolonged loss of communication | Trope |
| territory | Location |
| Geneva | Location |
| radar | Object |
| ABC1234 | CallSign |

⁴ onto.fel.cvut.cz/ontologies

5 Future work

The text analyzing tool for the aviation safety reports is aimed to be integrated into the reporting process workflow. For DBpedia spotlight, further work can be done regarding the disambiguation feature as well as taking context into consideration in the annotation process. Furthermore, domain-specific techniques will be taken into consideration. More artifacts can be declared and detected within the customized techniques. This will eventually raise the recall, precision and ultimately the F1 measure to score higher rates. In future research, we will focus on the relations detection between the concepts rather than only the concepts themselves. This will guarantee better understanding and analyzing for the reports.

Acknowledgements: This work was partially supported by grants No. TA04030465 Research and development of progressive methods for measuring aviation organization's safety performance of the Technology Agency of the Czech Republic, No.SGS16/229/OHK3/3T/13 Supporting ontological data quality in information systems of the Czech Technical University in Prague.

References

1. L. Tanguy, N. Tulechki, A. Urieli, E. Hermann, C. Rayna Natural language processing for aviation safety reports: from classification to interactive analysis Comput. Ind. (2015).
2. Timm Heuss, Bernhard Humm, Christian Henninger, Thomas Rippl, A comparison of NER tools w.r.t. a domain-specific vocabulary, Proceedings of the 10th International Conference on Semantic Systems, September 04-05, 2014, Leipzig, Germany.
3. Apache Stanbol - Working with Custom Vocabularies <https://stanbol.apache.org/docs/trunk/customvocabulary>. Last access 2016-06-29.
4. Apache Stanbol - Enhancement Engines. <https://stanbol.apache.org/docs/trunk/components/enhancer/engines/>. Last access 2016-06-29.
5. Joachim Daiber and Max Jakob and Chris Hokamp and Pablo N. Mendes, Improving Efficiency and Accuracy in Multilingual Entity Extraction, proceedings of the 9th International Conference on Semantic Systems (I-Semantics), 2013.
6. DBpedia Spotlight - Index.sh. <https://github.com/dbpedia-spotlight/dbpedia-spotlight/blob/master/bin/index.sh>. Last access 2016-06-29.
7. AIRCRAFT CALL SIGN FORMAT. https://ivao.aero/training/documentation/books/PP_ADC_Pilot_Callsign.pdf Last access 2016-06-29.

Hybridný prístup na klasifikáciu názorov

Katarína Simková, Martin Mikula, Kristína Machová

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky,
Technická univerzita v Košiciach, Slovenská republika

katarina.simkova@student.tuke.sk, {martin.mikula,
kristina.machova}@tuke.sk

Abstrakt. Na internete každý deň pribúda viac a viac textových dát. Tieto texty sú zaujímavým zdrojom informácií, ktoré môžu poslúžiť jednak ostatným ľuďom na internete, ako aj firmám zaoberajúcim sa predajom alebo marketingom. V prípade, že je týchto dát veľmi veľa, sú najlepším riešením ako ich analyzovať automatizované metódy. My sme sa v tejto práci zamerali na kombináciu 2 metód, slovníkového prístupu a Naivného Bayesovho klasifikátora. Táto kombinácia by mala odstrániť problémy oboch prístupov, keďže slovníkový prístup nevyhodnotí príspevky, ktoré neobsahujú slová zo slovníka a NBK zase potrebuje trénovaciu množinu, ktorú je náročné získať, hlavne pokiaľ sa jedná o novú doménu. Prezentovaný prístup dosiahol celkovú presnosť 57,63 %.

Typ príspevku: Výskumný príspevok

Kľúčové slová: slovníkový prístup, strojové učenie, Naivný Bayesov klasifikátor, klasifikácia názorov

1 Úvod

Analýza sentimentu je jednou z úloh spracovania prirodzeného jazyka (NLP – Natural Language Processing), ktorá zahŕňa získavanie a analýzu emócií autora na nejaké produkty, určovanie jeho názorov na politickú situáciu, alebo hodnotenie recenzií. Tieto hodnotenia sú väčšinou vyjadrované na internete prostredníctvom príspevkov na sociálnych sieťach alebo blogoch. V posledných rokoch výrazne narástlo množstvo názorov vyjadrovaných na webe a tieto názory sa stávajú stredobodom záujmu mnohých výskumníkov.

Sentiment analýza má svoje uplatnenie v rôznych oblastiach ako napr. v marketingu, kde pomocou sociálnych sietí a internetu sleduje reakcie zákazníkov na nové produkty a služby. Primárnou úlohou analýzy sentimentu je vyhľadať názor, identifikovať sentiment, ktorý tento názor vyjadruje a klasifikovať jeho polaritu. Sentiment vyjadruje ľudské pocity, emócie voči nejakému objektu. Najčastejšie delí do troch kategórií: pozitívny, neutrálny a negatívny[2].

2 Metódy používané na analýzu sentimentu

Metódy, ktoré sa používajú na analýzu sentimentu je možné rozdeliť do dvoch kategórií. Prvou sú metódy založené na strojovom učení a druhou sú metódy založené na slovníkovom prístupe. Okrem týchto dvoch existujú aj hybridné prístupy, ktoré tieto dve metódy kombinujú. Metódy strojového učenia používajú algoritmy strojového učenia, pomocou ktorých spracovávajú jednotlivé lingvistické vlastnosti analyzovaného textu. Medzi najčastejšie používané klasifikátory patria rozhodovacie stromy, lineárne klasifikátory, pravdepodobnostné klasifikátory alebo klasifikátory založené na pravidlách.

V práci Zhang a kol [6] autori porovnávali Naivný Bayesov klasifikátor a Metódu podporných vektorov na hodnoteniach reštaurácií. V práci taktiež rozoberali vplyv reprezentácie a veľkosti príznakového priestoru. Najlepšie presnosť dosiahol NBK používajúci 900-1100 atribútov a to 95,67%. K algoritmom strojového učenia môžeme priradiť aj metódy nekontrolovaného učenia prístupujú k riešeniu problému bez znalosti toho aký má byť výsledok. Pri nekontrolovanom učení sa príklady zhľukujú do zhľukov podľa nejakého kritéria, najčastejšie podobnosti. Medzi základné algoritmy patriace pod nekontrolované učenie patria zhľukovacie algoritmy [5].

2.1 Metódy založené na slovníkoch

V mnohých úlohách analýzy sentimentu sa využívajú slová, ktoré vyjadrujú naše názory a pocity. Tieto slová sa nazývajú tzv. názorové slová. Pozitívne slová sa používajú na vyjadrenie určitého požadovaného stavu, zatiaľ čo negatívne slová sa používajú na vyjadrenie nejakých nežiaducich stavov. Zoznam názorových slov sa nazýva slovník alebo lexikón. Takýto slovník sa následne používa na identifikáciu orientácie príspevkov. Lu a kol [3] určovali polaritu príspevkov pomocou násobenia prídavných mien a prísloviiek. Ich prístup dosiahol presnosť 71,7%. Možnostiam rozšírenia o využitie intenzifikátorov a negátorov je venovaná práca Kennedy a Inkpen [1]. Vo svojej práci porovnávali výsledky slovníka používajúceho intenzifikátory a negátory a slovníka bez nich. Slovník používajúci intenzifikátory a negátory dosiahol lepšie výsledky (priemerne 67,8%) ako pôvodný slovník (v priemere 66,5%).

3 Naivný Bayesov klasifikátor

V našej aplikácii sme sa rozhodli použiť Naivný Bayesov klasifikátor (NBK) z dôvodu jeho jednoduchosti a relatívne dobrým výsledkom. Je to jednoduchý klasifikátor založený na Bayesovskej teoréme, so silným dôrazom na nezávislosť atribútov. Veľmi často sa používa na jednoduchú klasifikáciu textov ako napr. vyhľadávanie spamu, filtrovanie mailov, kategorizáciu dokumentov, detekcia jazyka a analýzu sentimentu. Napriek silnému dôrazu na nezávislosť atribútov dosahuje Naivný Bayesov klasifikátor veľmi dobré výsledky aj v aplikáciách reálneho sveta.

V praxi sa používajú dva typy NBK [4]:

- Multinomiálny NBK - tento variant, odhaduje podmienenú pravdepodobnosť určitého slova k danej triede ako relatívnu početnosť termínu v dokumentoch, ktoré patria do kategórie C. Berie sa do úvahy počet výskytov termu v tréningových dokumentoch triedy C, vrátane viacnásobných výskytov.
- Bernoulliho NBK - generuje binárny ukazovateľ pre každý term zo slovníka, kde sa pridá hodnota 1, ak termín sa vyskytuje v dokumente a 0, ak nie. Tento model neberie sa do úvahy počet výskytov termov a berú do úvahy aj termy, ktoré sa v dokumente nevyskytli.

4 Návrh a implementácia hybridného modelu

Našou úlohou bolo vytvoriť hybridný model pre klasifikáciu názorov, ktorý by kombinoval slovníkový prístup a metódu strojového učenia. Bola vytvorená aplikácia, ktorá najskôr vytvorila tréningovú množinu pomocou slovníkového prístupu, následne na tejto tréningovej množine naučila NBK. Následne bol NBK použitý na testovaciu množinu. Aplikácia bola rozdelená na 3 časti:

Prvá časť aplikácie je zameraná na predspracovanie získaných dát. V tejto časti sa odstránia stop slová, odstráni sa diakritika, rozdelia sa slová a taktiež niektoré slová sa prevedú na základný tvar. Takto predspracované dáta nám potom slúžili ako vstup do ďalšej časti aplikácie.

Druhá časť je venovaná tréningu NBK, kde sa z jednotlivých príspevkov vytvorí zoznam slov, ku ktorým boli vyrátané apriórne pravdepodobnosti.

V predposlednej časti bol aplikovaný NBK na testovacie príspevky.

Proces klasifikácie bol rozdelený na dve časti. V prvej časti sa natrénuje NBK a v druhej je natrénovaný klasifikátor použitý na klasifikáciu testovacích prípadov. Na začiatku sa inicializujú prázdne matice, kde počet stĺpcov zodpovedá počtu slov v slovníku a počet riadkov zodpovedá počtu príspevkov v jednotlivých triedach. Na týchto maticiach prebieha učenie klasifikátora. Matíc je toľko, koľko tried budeme mať pri klasifikácii. Po načítaní sa prechádzajú príspevky a zisťuje sa, či je príspevok pozitívny alebo negatívny. Príspevky sa rozdelia na slová a pre každé jedno slovo sa zisťuje, či sa nachádza v zozname slov alebo nie. V prípade, ak sa slovo v zozname slov nenachádza na danom mieste ostáva hodnota 0, a v prípade ak sa slovo nachádza v zozname slov, tak zo slovníka sa vytiahne apriórna pravdepodobnosť pre dané slovo, a táto hodnota sa zapíše na pozíciu daného slova. Takýmto spôsobom sa vytvoria matice pre pozitívnu a pre negatívnu triedu. V procese klasifikácie sa následne musí inicializovať nulový vektor pre jednotlivý príspevok. Tento vektor bude taký dlhý, koľko slov sa nachádza v zozname slov. Po inicializovaní sa načíta príspevok, po jednom a rozdelí sa na slová. Neznámy príspevok sa prechádza, slovo po slove a každé slovo je dopytované vzhľadom na slovník. Ak sa dané slovo v slovníku vyskytuje, do vektora sa zapíše hodnota apriórnej pravdepodobnosti zo slovníka. Konečné hodnoty pravdepodobnosti sa následne počítajú pre pozitívnu ale aj pre negatívnu triedu zvlášť. Nakoniec je príspevok zaradený do triedy s vyššou pravdepodobnosťou.

5 Testovanie a vyhodnotenie aplikácie

Ako testovacie dáta sme vybrali príspevky vopred ohodnotené expertom. Príspevky predstavujú diskusiu na rôzne témy od politiky až po recenzie na spotrebnú elektroniku. Celý korpus sa skladá z 5242 príspevkov. Z tohto počtu bolo 4191 vyhodnotených pomocou slovníkového prístupu. Výsledky analýzy pomocou slovníkového prístupu boli ďalej použité ako trénovacia množina pre NBK. Zvyšných 1051 príspevkov, ktoré neboli vyhodnotené pomocou slovníkového prístupu bolo následne ako testovacia množina pre NBK. Z týchto 1051 príspevkov bolo 828 negatívnych a 223 pozitívnych komentárov. Výsledky presnosti a návratnosti sú zobrazené v Tab. 1.

Tab 3. Tabuľka obsahujúca presnosti na návratnosti NBK použitého v rámci hybridného prístupu.

| | Presnosť (%) | Návratnosť (%) |
|---------------------|--------------|----------------|
| pozitívne príspevky | 29,31 | 61,69 |
| negatívne príspevky | 85,94 | 61,17 |

Aplikácia dosiahla priemernú presnosť 57,63% a priemernú návratnosť 61,43%. F-miera pre pozitívne príspevky bola 0,39 a pre negatívne príspevky 0,71. Dosiahnuté výsledky mohli byť ovplyvnené niekoľkými faktormi. Prvým mohla byť nie úplne presná klasifikácia trénovacej množiny, ktorá bola vytvorená pomocou slovníkového prístupu. Ďalším faktorom mohla byť nevyváženosť príspevkov v testovacej množine.

6 Záver

V tejto práci bol popísaný hybridný prístup k analýze sentimentu. Bol vytvorený model, ktorý použil slovníkový prístup na vytvorenie trénovacej množiny pre metódu strojového učenia. Z metód strojového učenia bol vybraný Naivný Bayesov klasifikátor, ktorý bol po naučení použitý na príspevky, ktoré slovníkový prístup nedokázal vyhodnotiť. Priemerné výsledky okolo 57% resp. 61% nedosiahli naše očakávania, ale bolo by ich možné vylepšiť filtrovaním príspevkov, ktoré budú použité na trénovanie, alebo vyvážením testovacej množiny.

Podakovanie. Tento príspevok vznikol s podporou Vedeckej grantovej agentúry Ministerstva školstva, vedy, výskumu a športu Slovenskej republiky v rámci projektu č. 1/0493/16 „Metódy a modely pre analýzu prúdov dát“.

Literatúra

1. Kennedy, A and Inkpen, D.: Sentiment Classification of Movie and Product Reviews Using Contextual Valence Shifters. In: Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations (FINEXIN'05). Ottawa: 2005. s. 11-22.
2. Liu, B.: Sentiment Analysis and Opinion Mining [online] Available on: <<https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>>.

3. Lu, Y. a kol.: Exploring the Sentiment Strength of User Reviews. In: Proceedings of the 11th International Conference on Web-age Information Management (WAIM '10). Berlin: Springer-Verlag, 2010. s. 471-482. ISBN 978-3-642-14245-1.
4. Manning, Ch. D., a kol.: Introduction to Information Retrieval. University Press, 2008. 506 s. 1. edícia. ISBN 0521865719
5. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: A survey, Ain Shams University [online]. Available on: <<http://goo.gl/XWmv5E>>.
6. Zhang, Z., a kol.: Sentiment Classification of Internet Restaurant Reviews Written in Cantonese. In: Expert Systems with Applications: An International Journal. Roc. 38, c. 6 (2011). s. 7674-7682. ISSN 0957-4174.

Annotation:

Hybrid approach for opinion classification

There are new text data on Internet every day. These data contain a lot of interesting information, which can be useful for other people and also for companies, that deal with selling and marketing. In case, that we have huge amount of these data, it is useful to analyze them automatically. In our paper we focused on combination of 2 approaches for sentiment analysis, dictionary approach and Naïve Bayes classifier. This approach can solve the problem when the dictionary approach does not analyze any comments, because they do not contain any word from the dictionary. The second problem is that Naïve Bayes classifier needs training dataset, which can be difficult to obtain especially for new domain. The described approach achieved accuracy 57,63%.

Automatická anotácia a tvorba rečového korpusu prednášok TEDxSK a JumpSK

Ján Staš, Tomáš Koctúr, Peter Viszlay

Katedra elektroniky a multimediálnych telekomunikácií
Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach
Park Komenského 13, 041 20 Košice, Slovenská republika

{jan.stas, tomas.koctur, peter.viszlay}@tuke.sk

Abstrakt. Článok prezentuje nový korpus motivačných prednášok TEDxSK a JumpSK v slovenčine. Rečová databáza pozostáva z 220 prednášok v trvaní 58 hodín. Anotovaná množina rečových nahrávok bola vytvorená automaticky bez dohľadu pomocou akustickej segmentácie reči na báze analýzy hlavných komponentov a automatickej transkripcie pomocou dvoch komplementárnych systémov na rozpoznávanie reči. Pre potreby hodnotenia kvality automatického prepisu reči do textu bola vytvorená evaluačná množina 50 prednášok v trvaní 12 hodín s dodatočným manuálnym prepisom. Pomocou automatickej anotácie sme z rečového korpusu získali 21,26% nových rečových dát z celkovej doby trvania rečového korpusu pri zachovaní 9,44% miery chybovosti automatického prepisu vhodných na dotréňovanie, resp. adaptáciu pôvodného akustického modelu.

Typ príspevku: Príspevok o prebiehajúcom výskume

Kľúčové slová: automatické rozpoznávanie reči, automatická anotácia, rečový korpus, akustické modelovanie.

1 Úvod

Vývoj neustále presnejších systémov na automatický prepis reči do textu vyžaduje obrovské množstvo dát na estimáciu štatistických parametrov reči a jazyka, ktoré by pokryli čo možno najviac javov, ktoré sa v spontánnom rečovom prejave vyskytujú. Pri tvorbe robustných akustických modelov sa preto vyžaduje vybudovať foneticky bohatý a z pohľadu pohlavia vyvážený rečový korpus, ktorý by obsahoval rádovo stovky až tisíce hodín anotovaných rečových nahrávok. Vytvoriť takéto množstvo dát manuálne školenými pracovníkmi by zabralo neúmerne veľa času, ale aj finančných prostriedkov. Proces manuálnej anotácie je úmerný v priemere osem až desať násobku dobe trvania rečovej nahrávky [1]. Pri existencii určitého, aj keď len malého množstva manuálne anotovaných rečových dát je v súčasnosti možné pomocou najmodernejších prístupov a metód vybudovať komplexný systém na automatickú anotáciu a tvorbu nových rečových databáz, ktoré by mohli byť následne použité napr. pri reestimácii parametrov akustického modelu, resp. pri jeho adaptácii na hlasové charakteristiky hovoriacich.

Problémy pri vytváraní rozsiahlych rečových databáz možno vidieť aj na strane poskytovateľov zdrojových dát, s ich súhlasom. Z toho dôvodu sa hľadajú také zdroje, ktoré sú voľne dostupné širokej verejnosti. Jedným z takýchto zdrojov je aj databáza prednášok z konferencií TED (skr. z angl. *technology, entertainment, design*), ktoré sú organizované po celom svete a propagujú tzv. „myšlienky hodné šírenia“.

Vzhľadom na tematickú rôznorodosť prednášok a bohaté zastúpenie rečníkov, stali sa dobrým podkladom na tvorbu rečových databáz vo viacerých jazykoch. Jednou z najznámejších databáz je TED-LIUM [2], ktorá obsahuje celkovo 1495 automaticky anotovaných prednášok TEDx v anglickom jazyku. Miera chybovosti automatického prepisu (z angl. *word error rate*, skr. WER) dosahuje úroveň 17,40% v priemere. Systém na automatický prepis reči je založený na päť-prechodovom dekodovaní reči s postupnou adaptáciou akustických a jazykových modelov a reskóvaním hypotéz. Z najnovších rečových databáz možno spomenúť SI TEDx-UM [3], obsahujúcu 242 automaticky anotovaných prednášok TEDx v slovinskom jazyku. Miera chybovosti automatického prepisu WER bola v tomto prípade vyhodnotená pomocou systému na prepis spravodajských relácií BNSI a dosahovala úroveň až 50,70% v priemere.

Tento článok prezentuje novú rečovú databázu prednášok TEDxSK a JumpSK, anotovanú automaticky bez dohľadu pomocou dvoch komplementárnych systémov na rozpoznávanie reči v slovenčine s filtráciou hypotéz s minimálnym množstvom chýb. Databáza prepisov k rečovým nahrávkam bude zverejnená širokej verejnosti do konca roka 2016 na webovej stránke projektu Laboratória rečových a mobilných technológií¹.

2 Štruktúra rečového korpusu

Naším cieľom bolo vybudovať automaticky anotovanú rečovú databázu, obsahujúcu nahrávky v čo možno najlepšej kvalite s jedným, resp. dvoma rečníkmi na prednášku. Zdrojové dáta boli získané z kanálov TEDx Talks² a Jump Slovensko³ prostredníctvom internetovej služby YouTube. Zo zoznamu približne 300 motivačných prednášok z 10 podujatí zverejnených v rozmedzí rokov 2010 až 2016, boli manuálne vyradené všetky cudzojazyčné prednášky a nahrávky v nízkej kvalite. Z celkového množstva rečových nahrávok bolo vyselektovaných 220 prednášok v slovenskom jazyku v celkovom trvaní približne 58 hodín. Všetky audiovizuálne záznamy boli stiahnuté vo formáte H.264. Zachytená audiostopa bola zakódovaná vo zvukovom formáte MPEG AAC. Konverzia komprimovaného audia do formátu WAV (v 16-bit PCM mono audio) bola vykonaná pomocou nástroja SoX⁴. Všetky audiosúbory boli podvzorkované na 16 kHz, kvôli kompatibilite so systémom na automatické rozpoznávanie reči. Rečový korpus zahŕňa celkovo 227 unikátnych rečníkov, z toho 154 mužov a 73 žien. Zastúpenie ženských hlasov je približne 30% z celkovej doby trvania novovytvorenej rečovej databázy. Podrobný prehľad o zastúpení jednotlivých kategórií v novovytvorenom rečovom korpusu 220 motivačných prednášok TEDxSK a JumpSK je uvedený v Tab. 1.

¹ <http://nlp.web.tuke.sk>

² <https://www.youtube.com/user/TEDxTalks>

³ <https://www.youtube.com/user/jumpslovensko>

⁴ <http://sox.sourceforge.net/>

Tab. 1 Štruktúra rečového korpusu motivačných prednášok TEDxSK a JumpSK.

| názov podujatia | počet prednášok | počet rečníkov | z toho muži | z toho ženy | celkové trvanie | z toho muži | z toho ženy |
|----------------------|-----------------|----------------|-------------|-------------|-----------------|-----------------|-----------------|
| TEDx Bratislava | 57 | 61 | 42 | 19 | 13:03:55 | 09:02:35 | 04:01:20 |
| TEDx Kežmarok | 9 | 10 | 6 | 4 | 02:48:06 | 01:59:18 | 00:48:48 |
| TEDx Košice | 30 | 30 | 24 | 6 | 08:50:03 | 07:24:35 | 01:25:28 |
| TEDx Nitra | 14 | 14 | 12 | 2 | 04:13:37 | 03:33:07 | 00:40:30 |
| TEDx Prešov | 17 | 17 | 11 | 6 | 05:57:31 | 04:07:32 | 01:49:59 |
| TEDx Trenčín | 24 | 25 | 14 | 11 | 05:50:43 | 03:36:40 | 02:14:03 |
| TEDx Trnava | 9 | 9 | 6 | 3 | 02:21:53 | 01:42:20 | 00:39:33 |
| TEDxYouth Bratislava | 20 | 20 | 15 | 5 | 05:36:39 | 04:06:05 | 01:30:24 |
| TEDxYouth Žilina | 6 | 6 | 4 | 2 | 01:41:34 | 01:06:59 | 00:34:35 |
| Jump Slovensko | 34 | 35 | 20 | 15 | 07:27:35 | 04:12:44 | 03:14:51 |
| SPOLU | 220 | 227 | 154 | 73 | 57:51:36 | 40:51:55 | 16:59:41 |

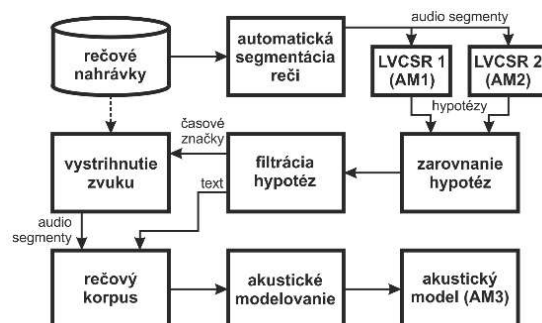
3 Automatická segmentácia a anotácia rečového korpusu

Princíp automatickej segmentácie a anotácie rečového korpusu prednášok pomocou dvoch komplementárnych systémov na rozpoznávanie reči je zobrazený na Obr. 1.

Automatická segmentácia reči pracuje na princípe segmentálnej analýzy hlavných komponentov (z angl. *principal component analysis*, skr. PCA) aplikovanej na časové vzorky mikrosegmentu reči, na ktorom sa vypočítajú a analyzujú jeho vlastné hodnoty. Pomocou nich sa determinuje charakter segmentu (rečová aktivita/tichý segment). Postupným vyhladzovaním a zásobníkovým akumulovaním elementárnych segmentov sa vytvoria kontinuálne segmenty reči bez tichých úsekov na základe preddefinovanej konfigurácie parametrov nami navrhnutého akustického segmentátora reči [4].

Inšpirovaní prácou [5] sme navrhli a vytvorili komplexný systém na automatickú anotáciu mohutných rečových databáz pracujúci bez dohľadu, ktorý je založený na komplementarite dvoch systémov na rozpoznávanie plynulej reči s veľkým slovníkom (z angl. *large vocabulary continuous speech recognition*, skr. LVCSR) v slovenčine [1][4]. Komplementarita spočívala v použití dvoch rôznych akustických modelov. Prvý z nich bol natrénovaný na databáze anotovaných nahrávok diktovanej reči v rozsahu 320 hodín, druhý model na databáze anotovanej spontánnej reči v rozsahu 330 hodín. Trigramový model slovenského jazyka bol obmedzený slovníkom 500k unikátnych slov a v procese dekódovania reči bol použitý voľne dostupný systém LVCSR Julius [6] s reskórovaním hypotéz pomocou mierne modifikovaného algoritmu ROVER [7].

Po automatickej transkripcii rečových nahrávok dochádza k porovnaniu a filtrácii výstupných hypotéz z oboch rozpoznávacích systémov LVCSR 1 a 2. V ďalšom kroku sú výstupné hypotézy zarovnané, pričom sa zohľadňuje okrem *minimálneho počtu zhodných slov* v zarovnaných hypotézach, tiež *maximálne časové odsadenie slov* od začiatku a konca rečovej nahrávky a *miera dôveryhodnosti správneho rozpoznania slov* (z angl. *confidence measure score*, skr. CMS). Výstupom sú potom krátke, automaticky anotované segmenty reči, ktoré prešli krokom filtrácie [1].



Obr. 1. Automatická segmentácia a transkripčia reči do textu pomocou dvoch komplementárnych systémov na automatické rozpoznávanie reči v slovenčine.

Experimentálne bolo tiež zistené, že minimálne časové odsadenie slov od začiatku a konca rečovej nahrávky je vhodné nastaviť na hodnotu 20 ms a počet zhodných slov v zarovnannej hypotéze by mal byť rovný minimálne trom slovám [1].

Tab. 2 Množstvo získaných dát po automatickej segmentácii a transkripcii rečového korpusu.

| množina | skutočné trvanie | trvanie po aut. segmentácii | nastavenie č. 1 ≈ 13,57% WER | nastavenie č. 2 ≈ 9,44% WER | nastavenie č. 3 ≈ 4,94% WER |
|-------------------------------------|------------------|-----------------------------|---------------------------------|--------------------------------|--------------------------------|
| množstvo získaných dát v [hh:mm:ss] | | | | | |
| eval | 12:26:07 | 11:50:37 | 05:39:30 | 02:47:35 | 00:39:43 |
| dev | 45:25:29 | 43:13:12 | 19:37:41 | 08:54:47 | 02:01:04 |
| eval + dev | 57:51:36 | 55:03:49 | 25:17:11 | 11:42:22 | 02:40:47 |
| množstvo získaných dát v [%] | | | | | |
| eval | | 95,24 | 47,78 | 23,58 | 5,59 |
| dev | | 95,15 | 45,41 | 20,62 | 4,67 |
| eval + dev | | 95,17 | 45,92 | 21,26 | 4,87 |

V prvom kroku budovania rečového korpusu prednášok TEDxSK a JumpSK sme rozdelili korpus na dve časti: evaluačnú (*eval*) a vývojovú (*dev*) časť. Evaluačná časť v rozsahu 12 hodín bola dodatočne manuálne anotovaná školenými pracovníkmi – anotátormi. Na tejto množine bola vyhodnotená účinnosť automatickej transkripcie reči do textu vo viacerých rôznych nastaveniach (pozri Tab. 2, nastavenie č. 1 až 3). Tieto hodnoty nastavenia systému boli zvolené s cieľom získať určitý objem anotovaných rečových dát s ohľadom na ich kvantitu (nastavenie č. 1), resp. ich kvalitu (nastavenie č. 3). Nastavenie č. 2 predstavuje kompromis medzi kvalitou a kvantitou automaticky anotovaných rečových dát. Následne boli tieto hodnoty nastavenia systému použité pri automatickej anotácii zvyšnej, vývojovej časti korpusu. Celkové množstvo získaných dát po automatickej anotácii rečového korpusu prednášok je zhrnuté v Tab. 2.

Z tabuľky možno pozorovať, že pri chybovosti automatického prepisu 13,57% WER sme získali približne 45,92% nových anotovaných rečových dát, ktoré môžu byť pou-

žité napr. na reestimáciu parametrov existujúceho akustického modelu, resp. jeho adaptáciu. Podobne, pri chybovosti 9,44% WER sme získali 21,26% nových dát a pri chybovosti 4,94% WER to bolo približne 4,87% dát z celkového množstva 58 hodín.

4 Záver

V tomto článku bol v krátkosti predstavený novovytvorený korpus prednášok TEDxSK a JumpSK. Anotovaná množina 220 rečových nahrávok bola vytvorená automaticky bez dohľadu pomocou systému na automatickú anotáciu a tvorbu rečových databáz, ktorý je založený na komplementaritete dvoch systémov na rozpoznávanie plynulej reči v slovenčine. Databáza prepisov k rečovým nahrávkam prednášok TEDxSK a JumpSK bude zverejnená širokej verejnosti do konca roka 2016 na webovej stránke projektu.

Podakovanie: Tento výskum bol realizovaný vďaka podpore Kultúrnej a edukačnej grantovej agentúry na základe Zmluvy č. 055TUKE-4/2016 a vďaka podpore Agentúry na podporu výskumu a vývoja na základe Zmluvy č. SK-HU-2013-0015 a realizáciou výskumného projektu APVV-15-0517, financovaných z prostriedkov MŠVVaŠ SR.

Literatúra

1. Kocút, T., Juhár, J., Vizslay, P., Staš, J., Lojka, M.: Unsupervised speech transcription and alignment based on two complementary ASR systems. In: Proc. of RADIOELEKTRONIKA 2016, Košice, Slovakia, (2016), pp. 358–362.
2. Rousseau, A., Deléglise, P., Esteve, Y.: TED-LIUM: An automatic speech recognition dedicated corpus. In: Proc. of LREC 2012, Istanbul, Turkey, (2012), pp. 125–129.
3. Žgank, A., Maučec, M.S., Verdonik, D.: The SI TEDx-UM speech database: A new Slovenian spoken language resource. In: Proc. of LREC 2016, Portorož, Slovenia, (2016).
4. Staš, J., Vizslay, P., Lojka, M., Kocút, T., Hládek, D., Kiktová, E., Pleva, M., Juhár, J.: Automatic subtitling system for transcription, archiving and indexing of Slovak audiovisual recordings. In: Proc. of LTC 2015, Poznań, Poland, (2015), pp. 186–191.
5. Naptali, W., Kawahara, T.: Automatic transcription of TED talks. In: Proc. of the 6th Spoken Document Processing Workshop, SDPWS 2012, Toyohashi, Japan, (2012), paper 16.
6. Lee, A., Kawahara, T., Shikano, K.: Julius – An open source real-time large vocabulary recognition engine. In: Proc. of EUROSPEECH 2001, Aalborg, Denmark, (2001), pp. 1691–1694.
7. Lojka, M., Juhár, J.: Hypothesis combination for Slovak dictation speech recognition. In: Proc. of the 56th Int. Symposium ELMAR 2014, Zadar, Croatia, (2014), pp. 43–46.

Annotation:

Automatic annotation and building of a speech corpus of TEDxSK and JumpSK talks

The paper presents a new Slovak spoken language resource built from TEDxSK and JumpSK lectures. The presented speech database consists of 220 lectures in total duration of 58 hours. Annotated speech corpus was generated automatically, in an unsupervised manner, by using acoustic speech segmentation based on a principal component analysis and automatic speech transcription using two complementary speech recognition systems. For evaluation of quality of

automatic transcription of speech, an evaluation set composed of 50 lectures, in total duration of 12 hours with manual transcription, has been created. Using automatic annotation of TEDxSK and JumpSK lectures, we have obtained 21,26% of a new speech data with 9,44% word error rate, suitable for re-training or adaptation of the original acoustic model.

otazkovac: A Question Generator for Slovak Stories

Marek Šuppa, Marek Nagy

Fakulta matematiky, fyziky a informatiky,
Univerzita Komenského v Bratislave
Mlynská dolina F1, 842 48 Bratislava, Slovenská republika

marek@suppa.sk, mnagy@fmph.uniba.sk

Abstract. In this work we describe *otazkovac*, a simple (web or command line) application capable of generating questions from specific types of Slovak sentences, provided it is fed with appropriate data. It is intended to be used as a submodule of *Multimedialna Citanka*, a web application thanks to which kids in the first three grades of Slovak primary schools learn how to read with some help of a computer. It uses a set of methods which are known to produce state of the art results in question generation problems on English text corpora. As part of this work we present a dataset based on stories from *Multimedialna Citanka* that can be used in further research on question generation from Slovak texts.

Contribution type: Application paper

Keywords: question generation, part of speech tagging, Slovak text corpora

1 Introduction

For the purpose of this work we will use the definition of Question Generation from [1] where it is defined as „the task of automatically generating questions from some form of input. The input could vary from information in a database to a deep semantic representation to raw text. Question Generation is viewed as a three-step process: content selection, selection of question type and question construction.“

The application presented in this work, *otazkovac*, tries to solve this task in a specific context of *Multimedialna Citanka*[2], which is a web application that helps children with improving their reading skills by analyzing a recording of their speech in real time and providing them with feedback on how accurately and how fast were they able to reproduce a given text (usually a short story). It also attempts to assess their comprehension skills by asking questions related to presented text. Since the creation of these questions by hand is a laborious task¹ a need for an automated solution arose. The aim of *otazkovac* is to fulfill this need.

¹ Especially considering the amount of texts in the database of *Multimedialna Citanka* and the fact that new texts are being continuously added.

While many complex and involved methods for Question Generation in educational contexts exist[3], given the intended use case of *otazkovac* it seems that a simple procedure of finding an appropriate sentence, detecting its type from the first few words and then syntactically transforming it into a question might be sufficient. A more detailed description of this procedure follows in the subsequent sections.

2 Question Generation

In order for *otazkovac* to find sentences that could potentially be turned into questions two stages are required: splitting text into sentences and detecting whether a sentence starts with a preposition. Both of these tasks can be performed by MorphoDiTa: Morphological Dictionary and Tagger[4], provided that it will get a pre-trained language model as an input. We were provided such a model by the Slovak National Corpus. While most of the publicly available Part of Speech (POS) taggers use the Penn Treebank POS tags, Slovak National Corpus uses a specific set of tags[5] that reflects the nature of Slovak language and provides more morphological information². Thanks to these tags, sentences which start with prepositions can easily be identified, as well as other words which belong to the “prepositional” part of the sentence.

An example of a tagged sentence looks as follows:

```
E---6-----u- - Po
AAis6----x----- - dobrom
SSis6----- - kúpeľi
R----- - sa
V-ms---cL-A-d---- - rozlúčil
E---7-----u- - s
SSis7----- - mesiačikom
O----- - a
SSfp7----- - hviezdíčkami
O----- - a
V-ms---cL-A-d---- - ľahol
R----- - si
V-----I-A-e---- - spať
Z - .
```

As we can see in the example, the first word's tag starts with *E*, which means that it is tagged as a preposition. Note that the next two words share the number (6) in their tags, in the same position as the first word. This is due to the fact that this position is used for the case of a word, and the tagger thinks that these words are all in the 6th Slovak case -- Locative. These three words can then be replaced with *Kedy*, the full stop can be replaced with a question mark and the result is a sentence that might constitute a fairly good question:

² Note that the tags discussed below come directly from the MorphoDiTa model and are slightly different from those described in [5]

Po dobrom kúpeli sa rozlúčil s mesiačikom a hviezdíčkami a ľahol si spať.
then becomes

Kedy sa rozlúčil s mesiačikom a hviezdíčkami a ľahol si spať?

3 Question Type Detection

Let us consider another example of a sentence with tags outlined for each word:

```
E---6-----u- - V
SSfs6----- - chate
V--p---aK-A-e---- - sme
R----- - sa
V-hp---aL-A-d---- - stretli
E---7-----u- - s
AAms7----x----- - ďalším
SSms7----- - poľovníkom
Z - .
```

As we can see the first two words in this example are of a type which is similar to the first three words in the first example. However, replacing them with *Kedy* does not seem like an option since in Slovak language: if the preposition *v* is followed by an object and this object itself is not a time reference (such as weekday or name of a month) this “prepositional clause” is most probably associated with a place, not a time. Therefore *Kde* would be way more appropriate in this case than *Kedy*.

Just from these two examples it is obvious that in order for *otazkovac* to create correct and relevant questions it needs to be able to detect what type of a question can be generated from a given sentence (if any). To do so we gathered a dataset of sentences that could possibly be transformed into questions as described above from all stories available on Multimedialna Citanka. MorphoDiTa models also provide the lemma along with a tag, and so we included this information in the dataset in order to make the detection more robust and prone to variation in natural language.

The dataset consists of 695 tagged sentences. Unfortunately, the premise from above does not hold in general (thanks to variability in natural language) and there are sentences like “O zvieratách sa dočítam v encyklopédii Svet zvierat” in the dataset that do not fall in either the *Kde* (marked **P** for **Place** in the dataset) or the *Kedy* (marked **T** for **Time** in the dataset) category. These sentences should be **Ignored** and are therefore marked **I** in the dataset. The final dataset contains 431 sentences in the Place category, 240 sentences in the Time category and 24 sentences in the Invalid category.

3.1 Feature Engineering

In order to use a machine learning algorithm for detection of question type it is necessary to represent its inputs as a set of features. A natural choice for features in a scenario like this would be n-grams over the list of lemmatized words. A slightly better

alternative might be to treat POS tags as words too. The motivation behind such a decision is that for instance a sentence of type **P** is more likely to have the preposition *na* followed by some sort of a noun represented by a tag *SSis6-----* rather than a specific noun itself. Since the dataset we have is very small in size, this setup should help us capture more variability in the data. One last improvement that might help even more would be the addition of concatenated bigrams from the beginning and the end of the list so that the words “v poslednej zákrute” would be represented by a feature vector similar to “v zákrute” since the middle word does not change the type.

3.2 Model Selection

There are multiple models to choose from when it comes to text classification. We might use a multinomial Naive Bayes classifier (NB) as a baseline, random forest classifier (RF) as an example of a model that tends not to overfit, and a SVM which is one of the recommended models when it comes to text classification on small datasets. All of the models were tested in combination with the features described above using 10-fold cross validation. The results are provided below:

Tab 4. The resulting testing accuracies of tested combinations of models and features

| | <i>NB</i> | <i>RF</i> | <i>SVM</i> |
|-------------|-----------|-----------|---------------|
| 2-3 normal | 87.36% | 83.90% | 85.90% |
| 2-4 normal | 88.21% | 85.76% | 86.19% |
| 1-4 special | 89.49% | 85.62% | 89.06% |
| 2-4 revers | 89.49% | 87.06% | 89.06% |
| 2-3 revers | 89.49% | 88.78% | 90.64% |

The numbers are the degrees of grams which were used (2-3 grams means that bigrams and trigrams were used), *normal* means setup with just lemmatized words, *special* is the setup described in the last paragraph of the section above and *revers* is the setup in which POS tags are treated as words. As it turns out our special handcrafted features are at best the same as POS tags with 4-grams. When we train the best model on the whole dataset we get the accuracy of 98.27 percent.

1.1 Error Analysis

It might be interesting to see in which cases did the model failed to predict the correct class. Let us consider the following sentence:

Na konci mesta si našiel lietajúcu motorku a ukradol ju.

Unfortunately in this case the MorphoDiTa model decided that the third word ‘*mesta*’ used a different case than the two before. This is not true, but given our premise described above (see section 2) only the first two lemmatized words were treated as features instead of the first three of them, which greatly affected the result.

Another example of a mistake made by model can be seen in the following sentence:
Na Havaj sa teším.

In this case the MorphoDiTa model thinks that Havaj is an abbreviation or a special entity of sorts, which is a situation our premise is not ready for. However, we can also see that in this case a completely different type of question could be generated (namely using *Kam*). It also needs to be noted that given the simplistic nature of the problem at hand (only questions starting with *Kedy* and *Kde* are generated), many sentences (such as for instance *Z chaty vyšla mačka*) will not be considered. This shows that there is potential for future improvement.

4 Conclusions and Future Work

We present an implementation of a simple method for generating specific questions from unstructured Slovak text. This method incorporates POS tagging as well as supervised learning of “question types” for sentences that start with a preposition. Although it is still considered to be work in progress, preliminary tests show that the questions it currently generates are sufficient in the context of Multimedialna Citanka.

While this work's focus is on just one possible way of generating questions thanks to simple sentence transformation³ this approach might be reusable in other contexts. With an appropriate training dataset and a MorphoDiTa model, it can also be used for another language. We would also like to note that this project is licensed under the GNU GPL license and can be obtained along with the aforementioned dataset from <https://github.com/mrshu/otazkovac>

Acknowledgment: We would like to thank Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences for providing us with a MorphoDiTa model of Slovak language.

References

1. Rus, Vasile, et al. "A detailed account of the first question generation shared task evaluation challenge." *Dialogue and Discourse* 3.2 (2012): 177-204.
2. Nagy, M. "Multimedia Reading Book - Utilization an XML Document Format and an Audio Signal Processing", *Slovko 2005*, Bratislava, ISBN 80-224-0895-6, p.141-146
3. Le, Nguyen-Thinh, Tomoko Kojiri, and Niels Pinkwart. "Automatic question generation for educational applications—The state of art." *Advanced Computational Methods for Knowledge Engineering*. Springer International Publishing, 2014. 325-338.
4. Straková, Jana, Milan Straka, and Jan Hajic. "Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition." *ACL (System Demonstrations)*. 2014.
5. Jazykovedný ústav Ľ. Štúra, SAV. "Morfológická anotácia textov Slovenského národného korpusu" http://korpus.juls.savba.sk/attachments/morpho_en/tagset-www.pdf Accessed: 2016-06-29

³ An example of such process can be the conversion of the sentence *Medvede žijú v lese* into *V lese žijú medvede*. Note that while the first sentence is not directly transformable into a question using the methods outlined in this paper, the second one is.

**Modelovanie používateľ'a,
personalizovaný web,
odporúčanie**

Projekt HIBER: hlbšie poznávanie správania sa človeka v digitálnom priestore

Mária Bieliková, Pavol Návrat, Jakub Šimko, Jozef Tvarožek, Michal Barla,
Róbert Móro, Eduard Kuric, Martin Labaj, Martin Konôpka

Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava, Slovenská republika
{meno.priezvisko}@stuba.sk

Abstrakt. V rámci začínajúceho projektu Informačné správanie sa človeka v digitálnom priestore sa zaoberáme interpretáciou záznamov činnosti používateľov v digitálnych priestoroch. Cieľom projektu je hlbšie porozumenie a tvorba metód a modelov automatickej interpretácie správania. Do úvahy berieme tradičné zdroje spätnej väzby ako aj vstupy z dosiaľ veľmi nevyužívaných senzorov ako okulografy vo výskumnom centre používateľského zážitku a interakcie na FIIT STU (uxi@fiit, <http://uxi.sk>).

Typ príspevku: Príspevok o prebiehajúcom výskume

Kľúčové slová: správanie používateľov, analýza dát, sledovanie pohľadu

1 Motivácia: porozumieť interakcii človeka s aplikáciami

Poznať správanie človeka v digitálnom priestore je dôležité pre úspech každej aplikácie, ktorú tento človek používa. Správanie, teda postupnosť akcií používateľa v aplikácii, nesie spolu s kontextom veľa informácií o tom, ako aplikáciu používa, či v nej dosahuje svoje ciele, či je príjemná a podobne.

Ak vieme správanie identifikovať (klasifikovať) automaticky, môže naň aplikácia reagovať (napríklad zabrániť jeho odchodu z aplikácie alebo vhodne odporučiť obsah). Automatické zisťovanie správania má okrem toho význam aj ex-post: uľahčuje analýzu aplikácie, užitočnú pri analýze použiteľnosti, marketingových stratégií a pod [2].

Surové dáta zachytávajúce správanie sú však nízkoúrovňové a je potrebné ich interpretovať. Elementárne používateľské akcie nachádzajúce sa v záznamoch len málokedy priamo hovoria o motívoch, cieľoch či pocitoch používateľov. Je však ťažkým problémom interpretovať záznamy o používaní do podoby symbolickej reprezentácie, ktorá by typy správania používateľov explicitne pomenovala. Časť prístupov sa snaží túto interpretáciu obísť riadením efektov adaptácie cez strojové učenie (napr. predikcia odchodu používateľa z aplikácie [1]). Takéto prístupy však často príliš závisia od konkrétnej domény a vyžadujú veľké vzorky trénovacích dát.

Interpretácii a hlbšiemu porozumeniu správania sa človeka v digitálnom priestore sa venujeme v začínajúcom projekte základného výskumu HIBER. Výskum rozvíjame predovšetkým na vybudovanej infraštruktúre Výskumného centra používateľského zážitku a interakcie, v laboratóriách UXI@FIIT, v ktorých máme k dispozícii viacero senzorových technológií umožňujúcich podrobné zaznamenávanie elementárnych akcií správania sa človeka pri práci s počítačom a teda aj pohybom v digitálnom priestore. Projekt nadviaže na viaceré existujúce výskumy zamerané na detekciu vzorov správania (snaha o zavádzanie a podvodné správanie) a typických stavov používateľov (emočné vybudenie, kognitívna záťaž).

2 Doterajší výskum správania sa používateľov v UXI@FIIT

UXI@FIIT funguje na fakulte viac než rok. Zameriavame sa na automatizované vyhodnocovanie použiteľnosti a podporou používateľských štúdií (vizualizácie dát, anotačné nástroje). Zároveň značnú časť úsilia venujeme výskumu metód automatického zisťovania stavu používateľov. Okrem analýzy “tradičných” zdrojov spätnej väzby (najmä sekvencie záznamov používania aplikácií) využívame analýzu dát zo špecializovaných senzorov, ktorými je laboratórium vybavené (ide najmä o sledovanie pohľadu a ďalej záznamy hĺbkových kamier, EEG a senzorov fyziológie ľudského tela). V rámci niekoľkých menších projektov sme skúmali možnosti detekcie všeobecných (na doméne čo najviac nezávislých) stavov, v ktorých sa používatelia môžu nachádzať a vedomost’ o ktorých môže do značnej miery prispieť k vhodnej adaptácii aplikácií.

Odhad emocionálneho vybudenie a kognitívnej záťaže pomocou merania rozšírenia zreničiek. Zmena priemeru zreničiek nastáva v prípade zvýšenia emocionálneho vybudenie alebo kognitívnej záťaže [3]. Použitie sledovania pohľadu na zisťovanie týchto stavov komplikuje reaktivnosť zreničky na svetelné podmienky. Tie sa menia aj zmenami na obrazovke počítača či zmenou cieľa pohľadu. Navrhli sme a overili personalizovaný model predikcie zmien zreničky z dôvodu zmeny svetelných podmienok, ktorý berie do úvahy miesto, na ktoré sa používateľ na obrazovke pozerá a vypočíta vnímanú svetelnosť, ktorú používateľ vníma a na základe toho predpovedá zmenu priemeru zreničky. Prvé výsledky boli prezentované na konferencii UMAP 2016 [5].

Meranie schopnosti vizuálneho hľadania. Jednou zo stratégií ako vysvetľovať správanie používateľov je modelovanie ich schopností. Napríklad schopnosť vizuálneho hľadania (schopnosť pohľadom lokalizovať prvok či informáciu v rozhraní) univerzálne ovplyvňuje výkonnosť používateľov pri vykonávaní úloh v aplikáciách, predovšetkým na webe. Vytvorili sme test vizuálneho hľadania, pri ktorom s pomocou sledovania pohľadu určíme úroveň tejto používateľovej schopnosti. Test je postavený na riešení umelých úloh vizuálneho hľadania ako aj úloh v reálnych rozhraniach webových stránok. V teste využívame okrem metriky reakčného času aj rôzne metriky sledovania pohľadu, najmä počet fixácií pri riešení úlohy.

Detekcia zavádzania pri vyplňaní dotazníkov. Zisťovanie nepoctivého správania používateľov má význam v mnohých scenároch, časté je pri vyplňaní dotazníkov. Uskutočnili sme experimenty, v ktorých sme nechali používateľov vyplňať osobnostný dotazník Big five, pričom raz mali participanti za úlohu vyplniť ho pravdivo a druhý

krát tak, aby sa čo najviac páčili potenciálnemu zamestnávateľovi. Celý priebeh vypĺňania sme zaznamenávali, vrátane sledovania pohľadu. Viaceré metriky, ako napríklad čas prvej fixácie na odpoveď a zmena priemeru zreničky, sa ukázali byť indikatívne pri rozlišovaní nepoctivého správania používateľov.

Porovnávanie sedení v rámci používateľských štúdií. V kontraste s predchádzajúcimi výskumami, pracovali sme aj na prístupe „zdola“, ktorý sa snaží dať sekvenciám používateľských akcií zmysel cez zhľukovanie (hľadanie opakujúcich sa vzorov správania). Prístup využíva učenie bez učiteľa, konkrétne obmedzený Boltzmannov stroj. Tejto forme neurónovej siete sú predstavované fragmenty sedení v podobe teplotných máp, v ktorých sieť dokáže nájsť vhodné abstrakcie. Abstrakcie sú následne vizualizované do prehľadnej schémy a pripravené na ďalšiu manuálnu inšpekciu [6].

3 Výzvy v projekte HIBER

Začínajúci projekt sa zameriava na výskum nových modelov a metód získavania a spracovania informácií dôležitých pre lepšie pochopenie informačného správania človeka v digitálnom priestore. Tieto modely a metódy otvárajú priestor k zefektívneniu činnosti človeka v digitálnom priestore najmä zmiernením dôsledkov problému kognitívneho preťaženia informáciami v rozsiahlych digitálnych priestoroch. Príkladom môže byť odporúčanie vhodných informačných zdrojov na základe automatického predpovedania informačného správania človeka v špecifickej doméne [4].

Medzi výzvy, ktorými sa zaoberáme v predkladanom projekte, patria:

1. *Limity kvantitatívneho uvažovania nad indikátormi implicitnej spätnej väzby.* Trendom existujúcich metód je analýza informačného správania ľudí na základe ľahko merateľných signálov a za ignorovania množstva ďalších faktorov, ktoré správanie ľudí môžu ovplyvniť. Možný smer je zapojenie hlbších, kvalitatívnych metód skúmania informačného správania človeka.
2. *Možnosti zapojenia nových signálov implicitnej spätnej väzby.* K „tradičným“ indikátorom ako kliky myšou, dopyty či rolovanie sa dnes pridávajú dosiaľ poriadne nepreskúmané indikátory ako sledovanie pohľadu, či fyziologické ukazovatele.
3. *Porozumenie správaniu používateľov.* Pozornosť pri modelovaní používateľa sa sústreďuje najmä na jeho ciele a na obsah digitálnych priestorov, potrebné je však aj rozumieť správaniu používateľov v procese dosahovania daných cieľov.
4. *Škálovateľnosť.* Pre všetky metódy a modely zároveň platí potreba ich škálovateľnosti a teda ich prispôbenie princípom veľkých dát a distribuovaného počítania.

Hlavným cieľom projektu je priniesť nové poznatky v informatike a informačných technológiách, najmä:

- Skúmať nové fenomény informačného správania človeka a priniesť nové poznatky spojené so správaním sa človeka v digitálnych priestoroch, v kontexte rôznych situácií a typov zariadení pre zber a poskytovanie informácií;
- Skúmať nové modely vystihujúce správanie ľudí v digitálnych priestoroch;

- Navrhnuť a verifikovať nové metódy získavania a analyzovania implicitnej spätnej väzby, predikcie informačného správania človeka a jeho využitia v zefektívňovaní aktivít ľudí v digitálnych priestoroch (najmä prostredníctvom personalizovanej navigácie a rôznych foriem vizualizácie digitálneho priestoru).

V projekte kladieme spolu s partnermi z Filozofickej fakulty UK v Bratislave dôraz na interdisciplinárnu analýzu dát (kvantitatívnymi a kvalitatívnymi prístupmi informatiky, informačnej vedy a psychológie). Po stránke technologickej sa orientujeme na nové zdroje spätnej väzby a technológie spracovania (prúdov) veľkých dát (potrebné na spracúvanie čoraz väčšieho množstva surových dát tečúcich zo zdrojov spätnej väzby). Záber projektu je pomerne široký a presahuje jednu oblasť poznania.

Rozbiehajúcim sa projektom chceme prispieť k metódam a modelom hlbšieho poznania správania používateľov v digitálnych priestoroch a k jeho automatickej interpretácii. Nadviazať tak chceme na dlhú výskumnú tradíciu v oblasti analýzy správania, modelovania používateľa a personalizácie. Do výskumu zároveň zapojíme prístrojovú a personálnu infraštruktúru laboratórií Centra používateľského zážitku a interakcie, a nadviažeme na tu uskutočňované existujúce projekty analýzy správania.

Podakovanie: Tento článok vznikol vďaka čiastočnej podpore Agentúry podpory vedy a výskumu v rámci projektu APVV-15-0508.

Literatúra

1. Kaššák, O., Kompan, M., Bieliková, M. (2016). Student Behavior in a Web-based Educational System: Exit Intent Prediction. In the Engineering Applications of Artificial Intelligence Journal, Elsevier, Vol.51, s. 136-149.
2. Paganelli, L. and Paternò, F. (2002). Intelligent Analysis of User Interactions with Web Applications. In Intelligent User Interfaces, IUI '02, USA, s. 111–118.
3. Zénon, A., Sidibé, M., Olivier, E. (2014). Pupil size variations correlate with physical effort perception. Frontiers in Behavioral Neuroscience, 8: 286.
4. Zheng, Y., Mobasher, B., Burke, R. D. (2013). The Role of Emotions in Context-aware Recommendation. Decisions@RecSys, ACM, s. 21-28.
5. Juhaniak, T., Hlavac, P., Moro, R., Simko, J., Bielikova, M. (2016). Pupillary Response: Removing Screen Luminosity Effects for Clearer Implicit Feedback. In UMAP 2016 Extended Proceedings. CEUR.
6. M. Barla, M. Šimek and M. Bieliková. Comparing Eye-tracking Data Using Machine Learning. In Journal of Eye Movement Research, Vol. 8, No. 4 (2015). Abstract. s. 192.

Annotation:

Project HIBER: deeper knowledge about user behaviour

We deal with the interpretation of user behaviour in the context of a starting project *Information behaviour of human in digital space*, and within the UXI@FIIT labs. The goal of the project is deeper understanding of user behavior as well as research of new methods and models of automated behavior interpretation. In this research, we especially take into account new sensoric sources of implicit feedback as well as traditional ones.

Discovering User Preferences in Gamification for Libraries: A Methodological Approach

Andrea Hrčková

Department of Library and Information Science
Faculty of Arts
Comenius University in Bratislava
Šafárikovo nám. 4, Bratislava, Slovak republic

`andrea.hrckova@uniba.sk`

Abstrakt. Gamification has still an untapped potential in library environment. The reason is that the successful application of gamification is not an easy task. Instead of imitating the applications consisting of points, charts of winners or badges, it is important to reflect the specific user motivations in the specific environment. With this aim, the empathy mapping method was applied in the design process, which was helpful mainly in terms of discovering user preferences of our target group in the fields of reading and gaming. The knowledge gained by this method was translated into the processes of web application using the user experience methods. It is believed that these methods can work as a bridge for the communication between the information science and computer science professionals and will help to accomplish the idea of a successful gamification application in libraries.

Contribution type: Work-in-progress paper

Keywords: library gamification, user experience methods, empathy map, customer journey

1 Introduction

A simple definition of gamification is “the application of game elements and game thinking in a non-game environment” [7]. The applications of gamification in a for-profit environment is often a successful method, how to attract new customers to the services in a playful manner. Therefore a right application of gamification may help to promote also libraries and reading. This research is unique, as gamification in libraries is a new domain [5], [6] and our user experience methods are also uniquely used in this area.

There are various methods of user testing in the field of user experience: eyetracking or mouse tracking combined with think – aloud methods, card sorting and wireframing, logs and web analytics, social media analysis, competitive intelligence, interviews, ethnography and surveys. The application of the first six methods is successful, when

testing interfaces or prototypes of services that are already existing and the aim of a stakeholder is just to improve the existing solutions. The disadvantage of an ethnography research is that it produces a lot of bias and is requiring in terms of time. The interviews or surveys are often not successful in finding innovative solutions, as users often don't know, what they need. It was necessary to think beyond these methods to create new, useful and usable solution.

The foundation of a good user experience is the knowledge about user, his problems, goals and needs. Deep understanding of users' behavior is crucial and since emotions influence behavior, either an understanding of users' emotions is important.

The main target group of users of gamification application – university students was set. The reason is that they still have time and potential for reading and also many players can be found amongst this group. Creating personas of our target group was the next step, as designers using personas created 80 percent more ergonomic design than designers that didn't use them [4].

2 Empathy map

For deep understanding of user behavior, empathy, or the identification with the feelings, thoughts, or attitudes of another is needed. Also therefore an empathy mapping method was chosen for our user research. An empathy map is a tool, helpful in synthesizing the observations about the users and in drawing out unexpected insights. Empathy maps vary in shapes and sizes, but there are basic elements common to each one [1]:

- Four quadrants broken into “Thinking,” “Seeing,” “Doing,” and “Feeling.”
- Sticky notes covering the quadrants (different color for different user)
- Additional boxes at the bottom of the quadrants: “Pains” and “Gains” (in some versions)

The first step in the process is the summarization of researcher's fieldnotes from user observations (sketches, audio/video files and photos) [1]. Instead of organizing the data in the quadrants by ourselves, the procedure included a direct brainstorming with users of our target group about their “Thinking,” “Seeing,” “Doing,” “Feeling” and “Pains” both in the fields of reading and playing games. Eight users that are playing and reading on a regular basis were selected for this qualitative user research. They were asked to do the exercise alone during the session by thinking about their favorite games and books, write it down on the sticky notes and put them to the appropriate quadrants on the empathy map. **The most important part of the statement was the "because" part,** where they were asked to explain their thoughts by a moderator.

After a team brainstorming, the specific themes and key concepts started to emerge and the primary needs of the users were identified. These themes were then sorted into categories and visually organized to the mind map on the brainstorming session. This approach formed a cohesive vision of the future user experience that was visualized in the form of customer journey afterwards.

According to the empathy mapping research results, the most important for both gamers and readers was the feeling of identification with the character and the immersion to the story. Young readers are used to connect and compare the information they read with their reality and the text make them think about values. In games, some more active motivations were additionally mentioned as the possibility to evolve, discover, cooperate or to kill the enemies. Both in reading and playing the feeling of thrill and conflict has to be present. The senselessness (missing objective), bad graphics or difficult text and cliché should be avoided. On the contrary, application should surprise users with the ability to build or to see something new. The gamers need to know the final objective so that they can strive to reach it and feel the adrenaline throughout the whole way.

3 Customer journey

A customer journey map is a visual interpretation of the overall **story from an individual's perspective** of his relationship with a product, service or organization over time and eventually across channels [2]. It allows to envisage interactions from the users' points of view, instead of taking an inside-out approach. The journeys can be used in both evaluation of current or future product/ prototype. They are useful to examine the present points of delight and pain points and uncover the opportunities for building a better user experience with the aim to fit the products or services into the users' lives. The basic components of customer journey are [2]:

- Timeline: a finite amount of time or variable phases
- Emotions: peaks and valleys illustrating emotions; or pain points and points of delight of user experience
- Touchpoints: customer actions and interactions with the application

There are various forms of customer journeys, of which the left to right timeline is mostly used. The other variations are circular or helical maps that can be supplemented by pictures or multimedia. There are some templates available online, the Game template by Uxpressia¹ was adjusted for our needs in our research.

4 Translating user needs to application attributes

A dominant need that was identified during empathy mapping session was the immersion to the story. Also according to Kelway [3] the feeling of total immersion from the interaction with a product is particularly prevalent in the gaming world. It results in the feelings of joy, satisfaction and escapism from reality. The manifestation in games may be in challenges that can be overcome. Sensory experiences that are in balance with cognitive engagement seem to provide the best experience [3].

The answers to question, how exactly a user can be enveloped by the gamification application were found in the further answers of respondents. Some of them mentioned

¹ <https://uxpressia.com/>

a conflict in game or in the fiction story, the others mentioned the ability of building something or destroying the enemy. The example of translation of one user need (conflict) together with the solution is depicted in the customer journey (table 1).

Tab 1. An example of translating user need to the system feature using the adjusted game template of customer journey (the use phase of the timeline)

| | USE |
|-------------------|--|
| USER NEEDS | Conflict |
| USER EXPERIENCE | Development / weakening of soul force in fights |
| TOUCHPOINTS | Fight on battleground |
| PROCESS | Players can invite the others to a duel, if strong enough. Players with same amount of points from different groups will meet on the battleground |
| PAIN POINTS | Death (player cannot be killed, just weakened) |
| PROBLEMS | Too easy / difficult questions |
| POINTS OF DELIGHT | Winning = destroying the force of a competitor, the raise of winner's force of spirit |

5 Conclusion

The methods, used in the research of gamification that shift the paradigm from self-centered to user-centered approach were explained briefly. Involving user preferences and problems in the first phases of information system design is crucial for a successful application. The gamification application will be created in cooperation of Department of Library and Information Science (Faculty of Arts, Comenius University in Bratislava), Faculty of Informatics and Information Technologies (Slovak Technical University in Bratislava) together with the volunteers (book reviewers and graphic designers) as a crowdsourcing project. Our methods were considered as a helpful communication bridge by the partners for writing the specification of the system and for its development. Still, further elaboration in the form of detailed wireframes is needed and is currently in the process. The resulting prototype will be evaluated with our target group and then implemented. The gamification application is planned to be deployed in Slovak libraries as a part of a particular library and information system. It is believed that the cooperation of social and computer sciences would be successful through the combination of their two different approaches.

Bibliography

1. *Empathy Map Method*. Stanford, Hasso Plattner Institute of Design, c2015. Available at: <https://dschool.stanford.edu/wp-content/themes/dschool/method-cards/empathy-map.pdf>
2. Grocki, M.: *How to Create a Customer Journey Map*. In: Uxmastery, 2014. Available at: <http://uxmastery.com/how-to-create-a-customer-journey-map/>
3. Kelway, J.: *UX design framework – Behaviour*. In: Userpathways, 2010. Available at: <http://userpathways.com/2010/02/ux-design-framework-behaviour/>
4. Long, F. *Real or Imaginary: The effectiveness of using personas in product design*. In: Irish Ergonomics Review, Proceedings of the IES Conference. Dublin: 2009. ISSN 1649-2102 - © Copyright 2009. Available also at: <http://www.frontend.com/the-effectiveness-of-using-personas-in-product-design.html>
5. Nicholson, S.: *Gamification in Libraries: A Word of Warning* [online]. Available at: <http://booksblog.infotoday.com/2012/12/gamification-in-libraries-a-word-of-warning/>
6. Spina, C.: *Gamification: Is it right for Your Library?* [online]. Available at: <http://www.aal-net.org/main-menu/Publications/spectrum/Archives/vol-17/No-6/gamification.pdf>
7. Werbach, K., Hunter, D.: *For the Win: How Game Thinking Can Revolutionize Your Business*. Philadelphia: Wharton Digital Press. ISBN 978-1-61363-023-5.

Pod'akovanie: Táto publikácia vznikla vďaka podpore projektov VEGA 1/0066/15 Modelovanie informačného prostredia digitálnej vedy a Fakultný grant FG15/2016.

Order Sensitive Measures of Preference Estimation Quality along Users

Michal Kopecky, Marta Vomlelova, Peter Vojtas

Faculty of Mathematics and Physics
Charles University
Malostranske namesti 25, Prague, Czech Republic

marta@ktiml|{{kopecky|vojtas}@ksi}.mff.cuni.cz

Contribution type: Work-in-progress paper

Keywords: data analytic, knowledge acquisition, user preference learning, recommender systems, order sensitive metrics, experiments

We consider user-item preference represented by a rating and deal with content based recommendation. We present our results on preference learning - models, methods, prototypes, data, metrics and experiments already published in [1, 2, 3, and 4] and add some material presented at EURO 2016 Preference Learning Stream [5].

We represent preferences on a set O of objects by rating (scoring) function $r: V \rightarrow [0;1]$, which assigns to every object $o \in V$ its overall preference score $r(o) \in [0;1]$. This score has pure comparative interpretation. We say an object o_1 is *more preferred* than object o_2 if $r(o_1) > r(o_2)$. Respectively, we consider $O \subseteq \Pi D_i$ the data cube (we freely switch between O and ΠD_i).

Assume further, we have a set of users U and for each user $u \in U$ the set $V^u \subseteq O$ of by him/her visited objects and corresponding observed rating $r^u: V^u \rightarrow [0; 1]$. Practically V^u is much smaller than O . Here we deal with offline testing (for online testing see [6]) and visited objects are divided $V^u = V^u_{\text{train}} \cup V^u_{\text{test}}$ to disjoint union of training and testing examples (with repeating cross validation). This implies, we have also $r^u_{\text{train}}: V^u_{\text{train}} \rightarrow [0; 1]$ and $r^u_{\text{test}}: V^u_{\text{test}} \rightarrow [0; 1]$. Our task is to find a recommendation in the form of a total rating $r^u_e: \Pi D_i \rightarrow [0; 1]$ such that r^u_e is a good estimation of r^u_{test} (in the sense of some metric, distance or order agreement (as r^u_e induces an ordering $<_e$)).

In [1] we described our team approach to RuleML 2015 Rule-based Recommender Systems for the Web of Data Challenge Track. The task was to estimate the top 5 movies for each user separately in a semantically enriched MovieLens 1M dataset measured by F-measure. We presented three methods. Surprisingly, the best recommendation was a domain specific method like "recommend for all users the same set of movies from Spielberg". Our main contributions were domain independent data mining methods tailored for top-k which combine second order logic data aggregations and transformations of metadata.

In [2] we introduced monotone preference models, i.e. models where r^u_e is a monotone composition of rankings on domains of explanatory attributes (possibly describing

user behavior, item content but also data aggregations). Target preference ordering of users on items is given by a preference indicator (e.g. purchase, rating). In this paper we focused on learning the (partial) order of vectors of rankings of user-item attribute values (without aggregation). We measure degree of agreement of comparable vectors with ordering given by preference indicators for each user. We are interested in distribution of this degree across users. We provide sets of experiments on user behavior data from an e-shop and on a subset of the semantically enriched Movie Lens 1M data.

In [3] we made a step further. We assume having explicit ratings with time-stamps from each user. We integrate three different movie data sets, trying to avoid features specific for single data and try to be more generic. We use several metrics which were not used so far in the recommender systems domain. Besides classical rating approximation with RMSE and ratio of order agreement we study new metrics for predicting Next-k and (at least) 1-hit at Next-k. Using these Next-k and 1-hit metrics we try to model display of our recommendation - we can display k objects and hope to achieve at least one hit. We trace performance of our methods and metrics also as a distribution along each single user. We define transparent and complicated users with respect to number of methods which achieved at least one hit. We provide results of experiments with several combinations of methods, data sets and metrics.

In [4] we wanted to test new methods and metrics. For this we designed a simulation. For instance first hit is the step in which a top-k item (from test set) appears in our estimation (the smaller the number the better). The ideal we would like to reach is to have for all users top-10 with first hit in estimated top-10. Unfortunately we are far from this. We depict results for parallel 1st hit measure, i.e. we consider test set (golden standard) ordering $<$ of items and estimation of ordering $<_e$, then the parallel 1st hit is the minimal position k in which $\text{top-k}(<) \cap \text{top-k}(<_e) \neq \emptyset$.

Second distinguished feature is that we measure quality of our prediction for each user separately. Results (of our method from [4]) are depicted in Figure 1 with box plot and with 5% and 95% percentile (visualization is cut at step 100).

We show results for generated data – two sorts of users (with either triangular shaped or bell shaped users' preference) and several types of data density (probability of explicit rating of an item in train set and in test set). We can see that in general triangular shaped users are easier to recommend than bell shaped users (e.g. for triangular users (in contrast with bell shaped) all medians are below 50, that is more than 50% of user has 1st parallel hit earlier than in the step 50 (if a web page depicts 10 results then there is a hit not later than on 5th page), only two groups with lower probability of rating are cut by step 100 (more than 25% of users has hit after step 100)). Second, we can see that users with higher probability of rating (rated 1 to 5 percent of items) are easier to recommend than those with lower probability (which rated 1 to 5 mille of items, i.e. 10 times less).

Nevertheless, we also see that comparing box plots generates a partial order – one group has better first quartile but median is worse. It is probably a task for business to say what is more relevant. We do not deal with implicit user behavior here. In general, our strategy is to interpret user's behavior as (fictitious) explicit rating, see also [6].

In [5] we considered user habits, how many of them are visiting second, third page of recommendation (we assume page is displaying 10 items).

Consider data from Table 1. 1st hit below 10, means hit on first visited page, 1st hit below 20 (above 10), means hit on second page, ... p_i - % of users with 1st hit $\leq i \cdot 10$, q_i - % of users visiting i^{th} page. We calculate aggregated success measure = $p_1 \cdot q_1 + p_2 \cdot q_2 + p_3 \cdot q_3$, results are depicted in Figure 2.

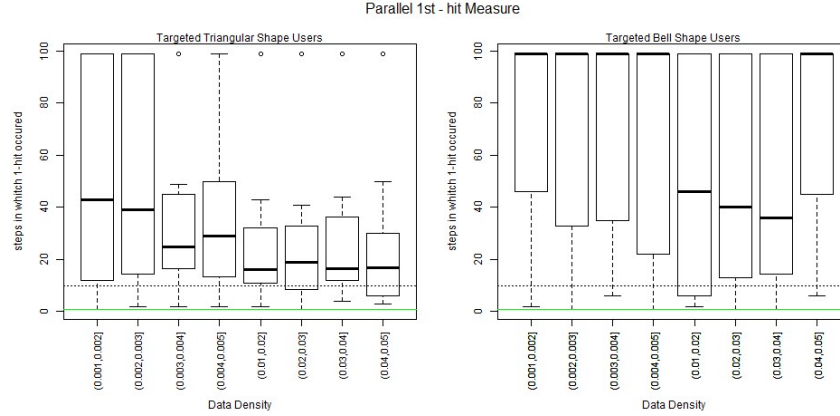


Fig. 4. Typical result for different user types, see also [4].

In [4] we introduced a new pivot based method. Pivot based method is better than remaining when measured in # in top-50 (1st hit). Results show that our data mining method is better than pivot when measured in RMSE

Tab. 1 Various user habits in visiting pages of recommended objects in ratio of users.

| | data 1 | data 2 | data 3 | data 4 | data 5 |
|----------------------|--------|--------|--------|--------|--------|
| 1 st page | 100 | 100 | 100 | 100 | 100 |
| 2 nd page | 9 | 86.8 | 92.04 | 86.72 | 51.4 |
| 3 rd page | 4.5 | 14.4 | 5.27 | 4.5 | 0 |

We made also several other comparisons. E.g. RMSE for random users – model-based vs. pivot based 3D methods compared show that our data mining model again gives better results than pivot based method.

We measured also sizes of intersection of test and estimated ordering at top-k. Our data mining was significantly less effective than pivot based method.

We report also on results of defended PhD, Master thesis in our seminar - especially on learning from implicit user behavior and/or online experiments.

A side product of our approach, the use pivots for collaborative filtering can contribute to cold start problem.

Future work is oriented in two directions. First is to improve results of this introductory investigation by further, more complex experiments both on artificial (more than

3D data) and both on real data. Second, we find interesting to study pivots based indexes on the space of observed ratings with respect to different metrics (distances, measures), especially order sensitive.

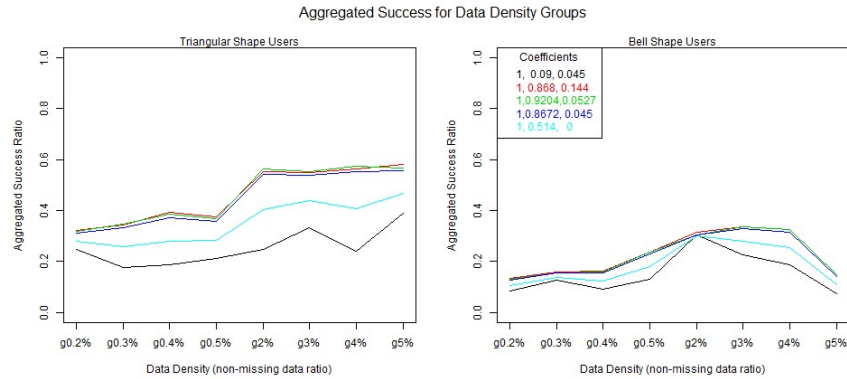


Fig. 5. Results from [4] recalculated with habits from Table 1, coefficients and colors correspond to columns in the table.

Acknowledgement. We announce partial support of Czech grant P46

References

1. Vomlelova, M., Kopecky, M., Vojtas, P. Transformation and aggregation preprocessing for top-k recommendation GAP rules induction. CEUR Proceedings Vol-1417 Rule-based Recommender Systems for the Web of Data, <http://ceur-ws.org/Vol-1417/paper18.pdf>
2. Kopecky, M., Peska, L., Vojtas, P., Vomlelova, M. Monotonization of User Preferences. In FQAS 2015, T. Andreassen et al Eds. Adv. Intell. Syst. Comp. 400, Springer 2016, 29-40
3. Vojtas, P., Kopecky, M., Vomlelova, M. Understanding Transparent and Complicated Users as Instances of Preference Learning for Recommender Systems, In Memics 2015, J. Kofron, T. Vojnar Eds. Lecture Notes in Computer Science 9548, Springer 2016, 23-34
4. Kopecky, M., Vomlelova, M., Vojtas, P. Basis functions as pivots in space of users preferences. In ADBIS (Short Papers and Workshops), M. Ivanovic et al Eds. New Trends in Databases and Information Systems Volume 637 of the series Communications in Computer and Information Science, Springer 2016: 45-53
5. P. Vojtas. Understanding Transparent and Complicated Users in Content Based Recommendation, presented at EURO 2016 Preference Learning Stream, Poznan 2016
6. L. Peska, P. Vojtas. Using implicit preference relations to improve recommender systems. J Data Semant, doi 10.1007/s13740-016-0061-8, Springer, published online 10 February 2016, <http://link.springer.com/journal/13740/onlineFirst/page/1>

DevACTs: Zber a vyhodnocovanie aktivít, úloh a zdrojového kódu vývojárov

Karol Rástočný, Martin Konôpka, Mária Bieliková, Pavol Návrat

Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava, Slovenská republika

{karol.rastocny, martin_konopka,
maria.bielikova, pavol.navrat}@stuba.sk

Abstrakt. Vývoj softvérových systémov je zložitý proces, ktorého sa zúčastňujú viacerí vývojári pracujúci na rôznych úlohách. Navyše každý vývojár má vlastné zvyky a postupy riešenia úloh, ktoré priamo ovplyvňujú vlastnosti vyvíjaného zdrojového kódu. Súčasné nástroje pre podporu vývoja softvéru ale sledujú aktivity vývojárov len v minimálnej miere, čím prichádzajú o významné dáta, ktoré umožňujú lepšie určiť vlastnosti vyvíjaného softvéru a zlepšiť sledovateľnosť samotného vývoja softvéru. Čiastočné riešenie tohto problému priniesla infraštruktúra vyvinutá v rámci projektu PerConIK. Táto infraštruktúra ale bola šitá na mieru výskumného projektu a neumožňovala jej distribúciu otvorenej komunite výskumníkov za účelom zberu a sprístupnenia dát. V tejto práci predstavujeme systém DevACTs vychádzajúci z infraštruktúry projektu PerConIK. Tento systém rieši problémy infraštruktúry projektu PerConIK a navyše je rozšíriteľný o ďalšie zdroje dát o aktivitách vývojárov so zdrojovým kódom, ako sledovanie pohľadu vývojárov alebo ich fyziológie.

Typ príspevku: Aplikačný príspevok

Kľúčové slová: sledovateľnosť vývoja softvéru, aktivity vývojárov, zdrojový kód

1 Motivácia

Analýza softvérových projektov je jedným zo základných prvkov manažmentu softvérových projektov, ktorej cieľom je identifikovať prečo nastali rôzne negatívne udalosti počas riešenia projektu [2]. Na základe tejto analýzy môžu projektoví manažéri upraviť existujúce procesy vo vývoji softvéru, prípadne definovať nové, aby zabránili vzniku identifikovaných negatívnych udalostí. Pri tejto analýze sú využívané najmä softvérové metriky vypočítané zo stabilizovaných verzií zdrojového kódu, výsledky testov, prípadne záznamy zo systémov správy úloh. Tieto dáta poskytujú iba informácie o udalostiach, ktoré nastali, ale poskytujú len minimálne informácie o procese ako tieto uda-

losti nastali. Chýbajúce informácie o procese vzniku môžu doplniť empirické softvérové metriky, ktoré sú vyhodnotené z aktivít vývojárov počas práce na softvérových artefaktoch.

Aj napriek schopnosti empirických softvérových metrik odpovedať na otázky „Prečo?“ (Prečo je v zdrojovom kóde toľko chýb? Prečo došlo k prekročeniu časového plánu?) sú tieto metriky využívané len v minimálnej miere. To je spôsobené najmä tromi problémami [3]:

- *Zber empirických dát je drahý na čas a zdroje* – veľa dát je zbieraných manuálne vývojármi;
- *Kvalita zbieraných empirických dát* – manuálne zbierané dát obsahujú veľa chýb a sú pomerne riedke;
- *Použiteľnosť empirických dát* – je definovaných len málo empirických softvérových metrik a málo nástrojov, ktoré umožňujú ich interpretovanie a ich analýzu.

Problém použiteľnosti empirických dát vyžaduje definovať nové empirické softvérové metriky. Pokusy o návrh empirických softvérových metrik [4, 5] boli aj v projekte PerConIK¹ [1]. Pri návrhu týchto metrik sa ale ukázali problémy zberu a kvality empirických dát. V projekte PerConIK sa podarilo pomocou navrhutej architektúry zozbierať pomerne veľké množstvo dát o aktivitách vývojárov a zdrojových kódach. Tieto dáta ale pochádzajú iba z približne 5 projektov od obmedzenej vzorky vývojárov – 15 programátorov v softvérovej firme a 20 študentov. Navyše zozbierané dáta sú pomerne riedke, keďže počas zberu sa vyskytlo viacero chýb, ktoré viedli k strate časti zbieraných informácií. Dôsledkom takýchto dát je, že prvotné hypotézy navrhovaných metrik často nebolo možné potvrdiť ani vyvrátiť.

2 Projekt DevACTs

Projekt DevACTs (Developer's Activity, Code and Tasks) je priamym pokračovaním projektu PerConIK, pričom jeho cieľom je vytvoriť zdieľanú infraštruktúru pre zber a analýzu empirických dát o vývoji softvéru. Infraštruktúra projektu DevACTs tak stavia na infraštruktúre projektu PerConIK, pričom sa snaží riešiť problémy, ktoré znemožňujú jej nasadenie v ďalších inštitúciách a komplikujú zber od vývojárov:

- *Decentralizovaná správa používateľov* – infraštruktúra projektu PerConIK nevyužíva centrálnu správu používateľov. Každý podsystém má vlastnú autorizáciu používateľov, pričom viaceré systémy sú silne zviazané, s konkrétnym nasadením protokolu LDAP. Táto decentralizácia prístupov k systémom znemožňuje nasadenie na iných inštitúciách a znemožňuje jednoznačné párovanie používateľov v zbieraných dátových množinách;
- *Anonymný zber aktivít vývojárov* – prostredie, v ktorom bola infraštruktúra nasadená umožňovalo len úplne anonymný zber aktivít od vývojárov. Ich mapovanie na

¹ <http://perconik.fiit.stuba.sk/>

zmeny v zdrojovom kóde tak muselo prebiehať na základe druhotných vlastností, čo viedlo k chybám v dátovej množine;

- *Zložitosť nasadenia* – webové služby projektu PerConIK si vyžadujú zložitú konfiguráciu prostredníctvom konfiguračných XML súborov, pričom jednotlivé konfigurácie prebiehali duplicitne a navzájom sa ovplyvňovali. Chybné nastavenie niektorej z hodnôt tak často viedlo k znefunkčneniu celej infraštruktúry. Navyše nasadenie monitorovacích nástrojov u vývojára pozostáva zo samostatných inštalácií viacerých nástrojov a chýba akákoľvek podpora pre automatické aktualizácie.

Na základe týchto identifikovaných problémov sme podrobili existujúcu implementáciu refaktorovaniu a odstráneniu známych chýb. Následne sme implementovali centrálny administratívny systém, ktorý vzájomne integruje ostatné podsystémy a zastáva tri hlavné úlohy:

- *Centralizácia konfigurácie* – každý podsystém je konfigurovaný prostredníctvom webového rozhrania, ktoré zabezpečuje kontrolu zadaných hodnôt a odstraňuje redundancie v konfiguráciách;
- *Diagnostika infraštruktúry* – centrálny systém zbiera záznamy o udalostiach v jednotlivých podsystémoch, čo umožňuje administrátorom jednoducho a včasne identifikovať problémy;
- *Správa používateľov* – centrálny systém je zodpovedný za autentifikáciu a autorizáciu používateľov. Ostatné podsystémy sú tak oslobodené od potreby ukladania citlivých údajov a riešenia základných prístupových práv. Taktiež administrátori majú možnosť jednoduchého nastavenia práv používateľov z jedného miesta, vďaka čomu je možné zabrániť neautorizovanému prístupu k dátam, ako aj zašumeniu zbieraných dát od neautorizovaných osôb.

Okrem vývoja centrálného systému, ktorý umožňuje jednoduché nasadenie infraštruktúry na nových inštitúciách sme sa sústredili aj na zlepšenie podpory zberu dát od vývojárov. K tomuto účelu sme prepracovali používateľské rozhranie klientskej aplikácie a implantovali systém aktualizácií a inštalátor, ktorý jednoducho prevedie vývojárov inštaláciou potrebných nástrojov.

3 **Závery a budúca práca**

Úpravy v infraštruktúre DevACTs nám dávajú možnosť zberu čistejších dát o aktivitách vývojárov a nasadenie tohto zberu nie len v ďalších inštitúciách, ale aj nasadenie monitorovacích nástrojov u nezávislých jednotlivcov, prípadne tímov, ktoré sa chcú podieľať na zbere dát o vývoji softvéru. Vďaka týmto dátam sa tak budeme môcť plnohodnotne sústrediť na výskum nových empirických softvérových metrík, ktoré umožnia lepšie porozumenie problémom v softvérových projektoch.

Do budúca sa v projekte DevACTs plánujeme zamerať na podporu vývojárov a projektových manažérov, pomocou ktorej budeme môcť poskytovať výstupy empirických softvérových metrík. Táto podpora je v súčasnosti v projekte reprezentovaná

prostredníctvom systémov CodeReview a CORD. Systém CodeReview poskytuje vývojárom priestor na posudzovanie zmien v zdrojových kódach, pokiaľ systém CORD sa zameriava na podporu manažmentu prostredníctvom vizualizácie artefaktov zdrojového kódu a ich vlastností. Napriek tomu, že oba systémy pracujú so zdrojovým kódom, boli vyvíjané rôznymi tímami, využívajú rôzne princípy poskytovania rovnakých, resp. podobných informácií. Tieto systémy plánujeme spojiť do jedného systému a minimalizovať tak zaťaženie vývojárov a projektových manažérov prácou s dvomi rôznymi systémami.

Pod'akovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektov VG 1/0752/14, VG 1/0646/15 a projektu v rámci OP Výskum a vývoj pre projekt: Výskum metód získavania, analýzy a personalizovaného poskytovania informácií a znalostí, ITMS: 26240220039, spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja.

Literatúra

1. Bieliková, M. et al.: Platform Independent Software Development Monitoring: Design of an Architecture. In: Geffert, V. et al. (eds.) SOFSEM 2014: Theory and Practice of Computer Science. pp. 126–137 Springer International Publishing, Cham (2014).
2. Buse, R.P.L., Zimmermann, T.: Information needs for software development analytics. In: 2012 34th International Conference on Software Engineering (ICSE). pp. 987–996 IEEE (2012).
3. Johnson, P.: You can't even ask them to push a button: Toward ubiquitous, developer-centric, empirical software engineering. In: The NSF Workshop for New Visions for Software Design and Productivity: Research and Applications. p. 5, Nashville (2001).
4. Konópka, M., Bieliková, M.: Software Developer Activity as a Source for Identifying Hidden Source Code Dependencies. In: Italiano, G.F. et al. (eds.) SOFSEM 2015: Theory and Practice of Computer Science. pp. 449–462 Springer Berlin Heidelberg (2015).
5. Kuric, E., Bieliková, M.: Estimation of student's programming expertise. In: Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement - ESEM '14. pp. 1–4 ACM Press, New York, New York, USA (2014).

Annotation:

DevACTs: Collecting and Evaluating Developers' Activities, Tasks and Source Code

Software system development is a complex process, to which many developers are involved. All these developers work on different tasks and they have different habits and ways how to solve their tasks. It directly influences characteristics of developed source code. Current software development tools support monitoring developers' activity, though minimally on their own, so that we miss important data needed for detailed analysis and evaluation of software projects. Infrastructure of the research project PerConIK partially solved this problem. However, this infrastructure was proposed only for specific research environment and it is not deployable for multiple teams of developers. In this paper we present infrastructure of the project DevACTs inspired by the project PerConIK. The DevACTs infrastructure solves problems of the original environment and it is extendable with new monitored events, e.g., gaze tracking or ECG.

Information behavior of researchers: contexts of digital scholarship

Jela Steinerová

Comenius University in Bratislava, Faculty of Arts,
Department of Library and Information Science
Gondova 2, 841 99 Bratislava, Slovak Republic

jela.steinerova@uniba.sk

Abstract. Information behavior of researchers in contexts of open science and digital scholarship is considered for understanding information needs and changing information infrastructures for scholarly communication. Our main research questions are: Which components build the conceptual framework for modeling information environment of digital scholarship? Which differences in information behavior of researchers of different disciplines can we identify? An analysis of models of digital scholarship is presented as the context of the research. A qualitative study of information behavior of 19 selected researchers is outlined based on semi-structured interviews. First results of content analyses are presented, including common general methodological approaches and different information interactions and publishing. In conclusion an ecological model of research information interactions is explained, composed of expertise factors, methodological factors, and open science factors.

Contribution type: Work-in-progress paper

Keywords: information behavior, researchers, digital scholarship, open science, research , information interactions

1 Introduction

Digital science is related to the transformation of creative scholarly communication and information processes into digital environments. Digital technological developments and digital data deluge have changed information behavior and information interactions. New types of documents and genres have emerged in digital environments, ranging from blogospheres to mobile digital libraries. Open science refers to research processes based on transparent information practices regarding methods, data, results and democratic access to knowledge and which allow broader public access to research results. Open science includes open access to scholarly literature, open data, open institutional repositories and electronic journals.

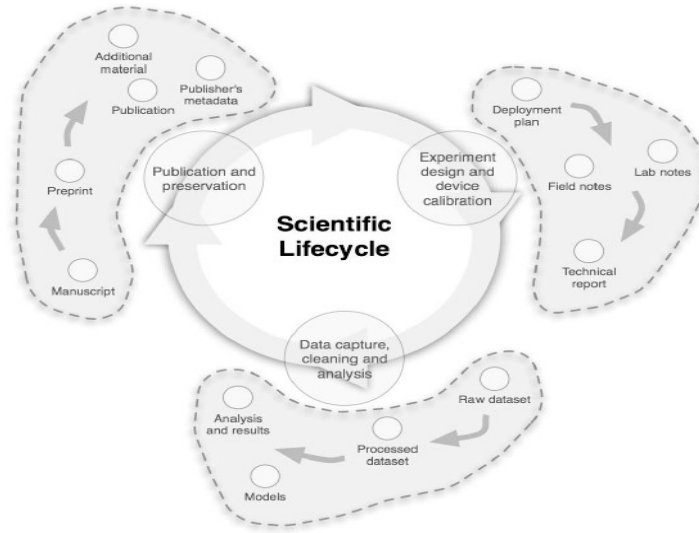


Fig. 2. Scientific life cycle example from the Center for Embedded Networked Sensing (Borgman, 2015, p. 265)

Chowdhury [5] presents a model of sustainable digital services based on sustainable information environment (Fig. 3).

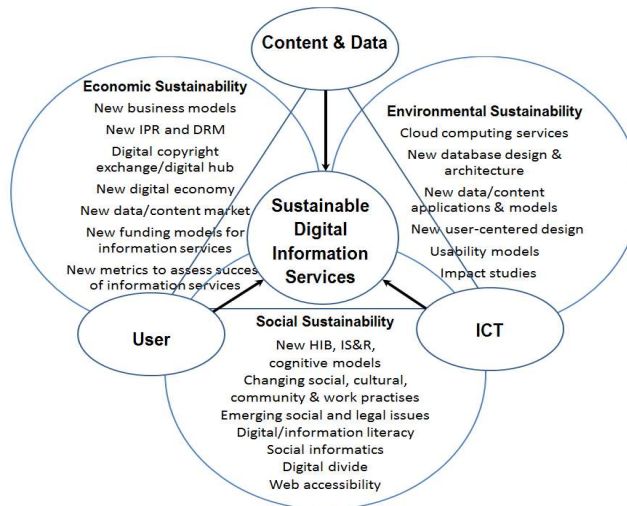


Fig. 3. Research issues and challenges in sustainable digital information services (Chowdhury, 2014, p. 195)

As a result of analyses of the models we can identify three basic components in digital and open scholarship: users and producers; knowledge infrastructure; and content, including artifacts and value-added services. They provide a common contextual background for conceptual modeling of information behavior of researchers ([6], [7], [8], [9]).

3 Information behavior of researchers: a qualitative study

In the framework of a research project on digital scholarship we carried out a qualitative study into the information behavior of 19 selected researchers in Slovakia. The main research question was focused on determination of domain differences with regard to information behavior of researchers and their perceptions of open science. We applied the methodology of semi-structured interviews.

A conceptual map was developed as a methodological tool for semi-structured interviews, content analyses and further conceptual modeling (Fig. 4).

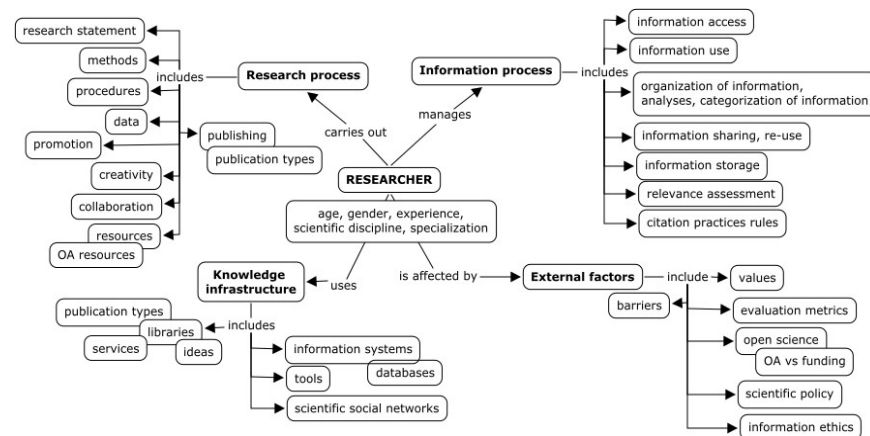


Fig. 4. Methodological design of the study (conceptual map)

The participants of the study included selected 19 researchers in sciences and medicine, humanities, social sciences and computer science in Slovakia. The selection criteria of subjects were based on the expertise and excellence in the domain, international networks, use of big data, advanced technologies and unique characteristics of the disciplines. The 19 respondents included 13 males (68,4 percent) and 6 females (31,6 percent), the average age was 54,4 and the average number of years of professional experience was 30 years. The representation of disciplines was composed of humanities (8, 39 percent), sciences and medicine (5, 28 percent), social sciences (4, 22 percent) and technical sciences (2, 11 percent). An average duration of an interview was 72 minutes. The interviews were carried out since October to December 2015 and since January to May 2016. The data were coded and frequencies of derived categories were interpreted.

Deeper semantic analyses are going on with the use of concept modelling and multiple analyses of different researchers in order to ensure the validity of results.

3.1 Results of first analyses

Relationships of scholarship with broader public, transparency of research processes and open access to data and publications were analysed. If we are to understand the information practices of open science in social, technological and community dimensions, we need to re-conceptualize the concept of research information interactions. Research information interactions can be determined as complex relationships between researchers and information environment. Following the ACRL Framework [10] we can determine research information literacy as the ability to understand and use information in order to carry out research in disciplines. However, not very much attention was paid to perceptions of open science and digital scholarship. That is why we analysed the data in relation to factors of open and digital scholarship.

These analyses point to common patterns and disciplinary differences in perceptions of knowledge infrastructure. Common patterns revealed common critical analytical information practices (information fluency). Practical experience and expertise is manifested by reliance on authoritative information sources and personal international expert networks. Open science factors were identified by researchers, especially promotion of results and open access. It is also connected with international participation, collaboration, peer networking, and information sharing (17 subjects). Technological determination, special methods and software tools were found especially with “big data” sciences, i.e. astrophysics, physics, genetics, archaeology, social sciences. In humanities, the tendency towards building digital collections and digital libraries was noted (e.g. archival system PamMap, Slavic languages atlas, archaeological photographic digital collections). Further open science factors included policies, evaluation of results, access to data and publishing. Awareness of researchers' social networks has been noted, including alternative metrics (altmetrics). Main differences emerged from domain-specific research objects, research statements, methodologies, procedures and data management. These differences are reflected in publishing activities (humanities: monographs, sciences: journals), communication, information use and culture of disciplines. Methodological modes of social sciences, humanities, sciences and technical sciences were identified.

4 An ecological framework of research information interactions

Based on results of analyses we developed an ecological framework of encapsulated research information interactions, composed of methodological factors, open science factors and expertise factors (Fig.5). Factors of open science (OS) include promotion, open access and participation. Several gaps with regard to open science were identified, namely the awareness of open access (OA) potential and promotion of research. The diagram represents intersections of processes which are relatively independent and in

mutual interactions create the holistic ecology of information interactions. These factors were derived from the content analysis of semi-structured interviews. The main insight is that methodological and OS factors are common to disciplines, while differences are based on expertise.

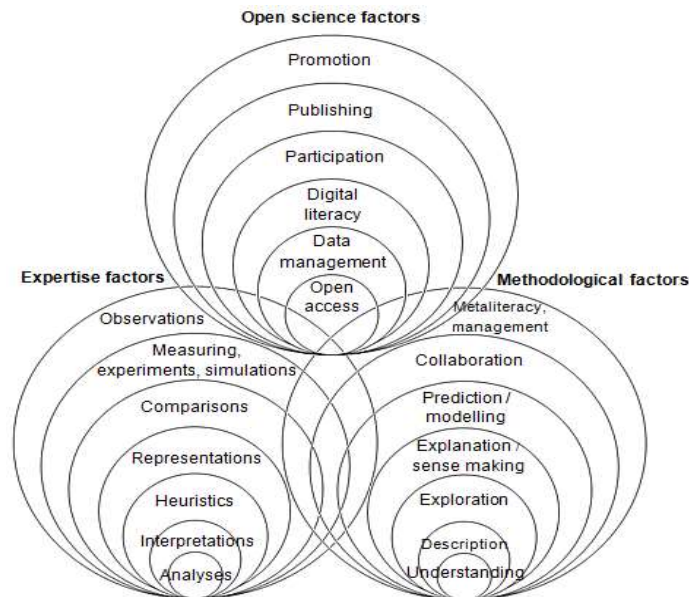


Fig. 5. The ecological framework of research information interactions

5 Conclusion

Models of digital scholarship and open science proved the need for deeper research into information needs of researchers. Based on this we developed a conceptual map which was used as a methodological tool for the qualitative study of information behavior of 19 researchers in Slovakia. Many differences among disciplines have been proved (e.g. retrospective nature and broad context of humanities, perspective nature and narrow context of sciences, specific methodologies, types of data and practices). Following the first analyses the ecological framework of encapsulated research information interactions was presented. We identified three groups of factors in information behavior of researchers, i.e. the expertise factors, methodological factors, and open science factors. Based on this we can determine research information literacy as understanding, sense making and knowledge discovery integrated with motivation and research interests. Our framework can be useful for development of knowledge infrastructures, including systems and services which actively support researchers in information practices, com-

munication and collaboration. Perceptions of open science can help reconstruct efficient partnerships between researchers, information professionals, librarians, research managers, institutions and research agencies.

Research information interactions can lead to changes in the workflow of the research and information processes and new models of digital environments for researchers. Support of information activities and creativity is needed in online genres and research communities of practice. Several components of digital environment (data, systems, tools, services) can contribute to new models of research and information processes. Further practical implications can be derived for value-added services and digital tools for researchers.

Acknowledgement: This work was developed with partial support of the project VEGA 1/0066/15 Modeling the information environment of digital science and the project supported by the Slovak Research and Development Agency under the contract No. APVV-15-0508 Human Information Behavior in the Digital Space

References

1. Hurd, J.: The Transformation of Scientific Communication: A Model for 2020. *JASIST*. Vol. 51. No.14, (2000), pp. 1279-1283.
2. Whitworth, B., Friedman, R.: Reinventing academic publishing online. Part II: A Sociotechnical vision. *First Monday*. Vol. 14, No.9, (Sept. 2009). Retrieved from: <http://firstmonday.org/ojs/index.php/fm/article/view/2642/2287>
3. Björk, B. Ch.: A Life-Cycle Model of the Scientific Communication Process. *Learned Publishing*. Vol.18, (2005), pp. 165–176.
4. Borgman, C. L.: *Big Data, Little Data, No Data. Scholarship in the Networked World*. Cambridge: MIT Press, (2015). 383 p.
5. Chowdhury, G. G.: *Sustainability of Scholarly Information*. London: Facet Publ., (2014). 231 p.
6. Ellis, D.: Ellis's Model of Information-Seeking behavior. In: *Theories of Information Behavior*. Medford, NJ: Information Today (2005), pp. 138-142.
7. Talja, S.: The Domain Analytic Approach to Scholars' Information Practices. In: *Theories of Information Behavior*. Medford, NJ: Information Today (2005), 123-127.
8. Fidel, R.: *Human Information Interaction. An Ecological Approach to Information Behavior*. Cambridge: MIT Press (2012). 348 p.
9. Case, D. O.: *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior*. 3rd. ed. Bingley: Emerald (2012). 491 p.
10. ACRL Framework Information Literacy for Higher Education. ACRL 2016. Board of Directors. 1-19 (Febr. 2, 2016). Retrieved from: http://www.ala.org/acrl/sites/ala.org/acrl/files/content/issues/infolit/Framework_ILHE.pdf.

Prezentácia personalizovaných odporúčaní v prostredí webu

Martin Svrček, Michal Kompan

Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava, Slovenská republika

{meno.priezvisko}@stuba.sk

Abstrakt. V súčasnej dobe sú personalizované odporúčania veľmi populárne a čoraz viac využívané. Avšak jeden zo základných problémov v tomto kontexte je nedôvera používateľov v odporúčacie systémy. Považujú ich za narušenie ich súkromia z dôvodu využívania pomerne osobných informácií. Preto je dôležité, aby boli odporúčania pre používateľov transparentné a zrozumiteľné. Pre riešenie týchto problémov sme sa rozhodli zamerať na oblasť prezentácie výsledkov odporúčaní. Konkrétne sme sa zaoberali vysvetľovaním jednotlivých položiek odporúčaní koncovému používateľovi. V tomto kontexte sme vytvorili odporúčací systém ExplORE, ktorý využíva kolaboratívne filtrovanie ako štandardnú techniku odporúčaní. V rámci tohto systému sme navrhli a implementovali hybridnú metódu personalizovaného vysvetľovania. Táto metóda je nezávislá od techniky odporúčaní a kombinuje tri základné prístupy k vysvetľovaniu s cieľom poskytnúť používateľovi vhodný typ personalizovaného vysvetlenia. Jednotlivými prístupmi sú vysvetľovanie založené na podobných používateľoch, vysvetľovanie založené na obsahu a vysvetľovanie založené na znalostiach o používateľovi.

Typ príspevku: Výskumný príspevok

Kľúčové slová: personalizované odporúčanie, vysvetľovanie odporúčaní, spravodajská doména

1 Úvod

Závažným problémom, ktorý vo výraznej miere bráni väčšiemu rozšíreniu a uplatneniu odporúčaní je často nedôvera používateľov. Táto nedôvera pramení aj z toho, že odporúčania využívajú pomerne osobné informácie o používateľoch. Tieto informácie sa môžu týkať ich znalostí, správania sa na stránke alebo určitých sociálnych charakteristík (priatelia, komunity, a pod.). Tieto systémy teda produkujú odporúčania bez nejakého bližšieho vysvetlenia alebo interpretácie dôvodu prečo je dané odporúčanie pre používateľa vhodné.

Zaujímavým prístupom k riešeniu takéhoto problému je rôzna forma prezentácie odporúčaní. Takáto prezentácia môže súvisieť napríklad s:

- Umiestnením a vizualizáciou odporúčaní tak aby zaujali používateľa, boli viditeľné a použiteľné.
- Vysvetľovaním samotných odporúčaní a procesu ich generovania, ktoré sa snažia priblížiť odporúčania používateľovi a tým riešiť problém s ich nedôverou.

Vysvetlenia sú orientované priamo na používateľov a snažia sa im opísať dôvody prečo by dané odporúčania mohli byť pre nich nápomocné [3]. Takéto vysvetlenia však neriešia a ani nemajú riešiť problémy s nesprávnosťou odporúčaní. Ich cieľom je iba podať odporúčanie vo forme, ktorá bude používateľom bližšia.

Z dôvodu pretrvávajúcich problémov s prijímaním odporúčaní samotnými používateľmi je našou snahou zamerať sa na lepšie podanie jednotlivých objektov odporúčania koncovému používateľovi. Chceme aby odporúčania nepôsobili odstrašujúco, ale aby boli naopak prijímané ako pomoc alebo podpora. Tomu sme prispôbili nielen metódu vysvetlenia ale aj samotnú prezentáciu alebo vizualizáciu odporúčaní.

V rámci snahy zaujať používateľa je teda na jednej strane dôležité mať kvalitný algoritmus pre odporúčania avšak rovnako dôležité je aj tieto odporúčania prezentovať čo najlepším spôsobom [1]. Niektoré výskumy dokonca ukazujú, že samotná prezentácia je v určitých prípadoch dôležitejšia ako technika odporúčaní [4]. V oboch prípadoch však ide o snahu vytvoriť užitočný systém z hľadiska používateľa. V tomto prípade sú základnými bodmi pre dosiahnutie užitočného odporúčacieho systému [2]:

- Potreba vyvolania dôvery v používateľoch
- Transparentnosť systému vzhľadom na používateľa
- Doplňujúce informácie o odporúčaní (obrázky, hodnotenia, a pod.)

2 Metóda hybridného vysvetľovania

Hlavnú myšlienku navrhutej metódy personalizovaného vysvetľovania predstavuje prístup generovania vysvetlení s ohľadom na preferencie používateľov. Každý používateľ teda má k dispozícii vysvetlenie prispôbené tak aby ho čo najviac zaujali.

Navrhnutá metóda je nezávislá od techniky akou bol odporúčený daný článok. Na druhej strane však vychádzame z prístupov k odporúčaniam pri generovaní vysvetlení. To znamená, že jednotlivé vysvetlenia sú ako keby položky odporúčania. Takýmto spôsobom na základe informácií o odporúčanom článku a informácií o používateľovi vygenerujeme alebo odporučíme vysvetlenie, ktoré bude vhodné a zaujímavé.

Výsledná metóda predstavuje určitý typ hybridného personalizovaného vysvetľovania. Hybridné preto, lebo kombinuje viaceré prístupy tak aby sme dosiahli optimálny výsledok. Týmito prístupmi sú:

- Vysvetľovanie založené na podobných používateľoch
- Vysvetľovanie založené na obsahu článkov
- Vysvetľovanie založené na znalostiach o používateľovi

Personalizované zase z toho dôvodu, že každý používateľ má v rámci daného prístupu zobrazenú konkrétnu informáciu, z ktorej bolo vysvetľovanie odvodené:

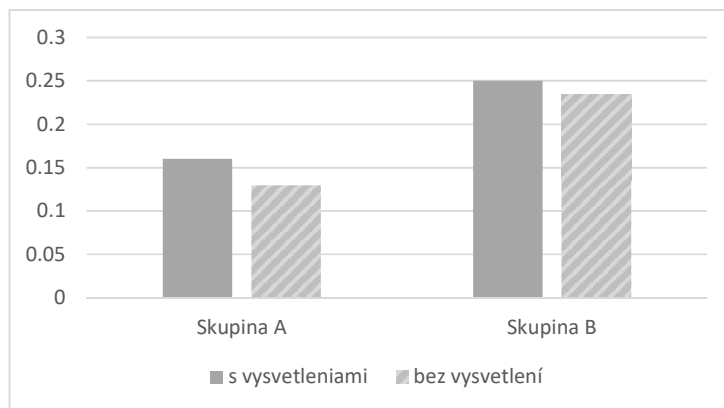
- Vysvetľovanie založené na podobných používateľoch – v tomto prípade bude zobrazený konkrétny používateľ využitý pre vysvetlenie.
- Vysvetľovanie založené na obsahu článkov – v tomto prípade bude zobrazený konkrétny článok využitý pre vysvetlenie.
- Vysvetľovanie založené na znalostiach o používateľovi – v tomto prípade bude zobrazená konkrétna znalosť využitá pre vysvetlenie.

V kontexte hybridného vysvetľovania je potreba kombinácie týchto prístupov. Takto poskytneme typ vysvetlenia, ktorý sa hodí pre daného používateľa. Pri tejto metóde najskôr prostredníctvom monitorovania preferencií používateľa nájdeme taký typ vysvetľovania, ktorý je vhodný v kontexte vlastností daného článku a ktorý je zároveň vhodný aj pre daného používateľa. Následne sa určí, ktorý typ vysvetlenia sa hodí pre konkrétného používateľa.

3 Overenie a záver

Navrhnutú metódu vysvetľovania sme implementovali ako súčasť systému ExplORe. V tomto systéme prebiehal dlhodobý experiment so simulovaním podmienok reálneho média s novinovými článkami. V rámci tohto experimentu sme zbierali údaje o aktivitách používateľov v systéme. Dĺžka trvania tohto experimentu bola 18 dní. Experimentu sa celkovo zúčastnilo 17 ľudí. Celkovo 13 z nich tvorili študenti vysokoškolského štúdia. Až 15 účastníkov bolo vo veku 20-30 rokov. Experiment bol rozdelený na dve časti kedy časť používateľov vysvetlenia k dispozícii nemala a druhej skupine vysvetlenia ponúknuté boli.

Vysvetlenia mali v oboch skupinách používateľov pozitívny vplyv na mieru ich aktivity v systéme a teda aj logicky na samotnú presnosť odporúčaní, ktoré im boli generované rovnakou metódou. Na Obrázku 1 je jasne vidieť, že oboch prípadoch je zaznamenaný nárast počtu klikov na články s vysvetleniami v kontexte presnosti.



Obr. 1. Presnosť klikov medzi skupinami pre články bez a s vysvetleniami.

Hybridná metóda personalizovaného vysvetľovania dosiahla pomerne pozitívne výsledky vo viacerých oblastiach nášho výskumu. V tomto kontexte preto vidíme aj pomerne veľký potenciál ďalšieho výskumu v tejto oblasti. V práci sme sa snažili prispôbiť vysvetlenie konkrétnemu používateľovi. Avšak rovnako zaujímavé je pokúsiť sa prispôbiť vysvetlenie aj konkrétnemu novinovému článku. Vysvetľovanie pre články súvisí so snahou nájdania spôsobu ako vysvetľovať určitý typ článkov. Tu sa môže ukázať, že pre určité oblasti alebo témy je vhodný konkrétny typ vysvetlenia.

Pod'akovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektov...

Literatúra

1. Ricci, F., Rokach, L., Shapira, B.: Introduction to recommender systems handbook. Springer US, 2011.
2. Swearingen, K., Sinha, R.: Beyond algorithms: An HCI perspective on recommender systems. In: ACM SIGIR 2001 Workshop on Recommender Systems. (2001), pp. 1-11.
3. Tintarev, N., Masthoff, J.: Evaluating the effectiveness of explanations for recommender systems. User Modeling and User-Adapted Interaction, (2012), 22.4-5, pp. 399-439.
4. Ziegler, C-N., McNee, S. M., Konstan, J. A., Lausen, G.: Improving recommendation lists through topic diversification. In: Proc. of the 14th int. Con. on World Wide Web. ACM, (2005), pp. 22-32.

Annotation:

Presentation of personalized recommendations

Nowadays, personalized recommendations are widely used and very popular. However, one of the basic problems is the distrust of users of recommendation systems. They consider them as intrusion of their privacy. Therefore, it is important to make recommendations transparent and understandable to users. To solve these problems, we decided to focus on the area of presentation of results recommendations. Specifically, we focused on explanation of each recommendation item to the end user. In this context, we have created a recommendation system EXPLORE, that uses collaborative filtering as a standard recommendation technique. Under this system, we have designed and implemented our hybrid method of personalized explanation of recommendations. This method is independent of recommendation technique and combines three basic approaches to explanation, in order to provide appropriate type of personalized explanations to the end user. Three basic approaches are explanation based on similar users, explanation based on content, explanation based on knowledge about user.

**Modelovanie informácií
a znalostí,
reprezentácia sémantiky**

OpenPonk - platforma pro konceptuální modelování ve výuce, vědě a praxi

Jan Blizničenko, Peter Uhnák, Robert Pergl

Katedra softwarového inženýrství
Fakulta informačních technologií
České vysoké učení technické v Praze
Thákurova 9, 160 00 Praha 6, Česká republika

{bliznjan, uhnakpet, robert.pergl}@fit.cvut.cz

Abstrakt. OpenPonk (<https://openponk.github.io>), původně známý jako Dyna-CASE, je vznikající open-source softwarová platforma pro konceptuální modelování. Cílem je podpora intuitivní tvorby, úprav, ověření a využití modelů. Podporuje také doménově-specifické jazyky (DSL). V současnosti OpenPonk umožňuje různě pokročilé práce s konečnými automaty, Petriho sítěmi, BORM ORD, DEMO, OntoUML a s diagramy tříd UML. Motivací je nabídnout otevřenou a snadno rozšiřitelnou platformu pro implementaci modelovacích notací a algoritmů. Cílovou komunitu představují vyučující, výzkumníci a odborníci z praxe.

V porovnání s dalšími současnými řešeními, které jsou zpravidla založeny na Java/Eclipse/EMF/GMF, naše řešení využívá čistě objektově orientovanou technologii Pharo/Roassal, kterou je výrazně jednodušší si osvojit pouhým pozorováním, zkoušením a následným napodobením. Navíc jde o živý systém, ve kterém je možná interakce uživatele s modely přímo během vývoje.

Popisujeme také projekt založený na OpenPonk pro francouzskou instituci CIRAD.

Typ příspěvku: Aplikační příspěvek, Příspěvek o probíhajícím výzkumu

Klíčová slova: CASE, CABE, Pharo, Smalltalk, Konceptuální modelování

1 Úvod – myšlenky a cíle OpenPonk

OpenPonk je platforma pro konceptuální modelování, která vzniká v živém prostředí Pharo[2].

Pro potřeby našeho výzkumu týkajícího se ontologií a konceptuálního modelování potřebujeme nástroj, který podporuje nejen práci se současnými notacemi, ale umožní i uživatelsky přívětivou možnost implementace nových notací a jejich následné zkoumání a využití. Krom toho se zajímáme o podporu notací a modelů vytvořených na míru ve firemním prostředí. Nyní jsou implementována různě pokročilá využití pro práci s konečnými automaty, Petriho sítěmi, BORM ORD[11], DEMO, OntoUML a s diagramy tříd UML.

Cílem našeho nástroje není přímá konkurence současným komerčním nástrojům jako je Enterprise Architect[14] nebo MetaEdit+[8], ale zaměřujeme se spíše na použití ve vědě a výzkumu, kde je výhodou nezávislost na platformě, otevřený kód a snadná rozšiřitelnost nejen o nové modely, ale i o další funkce. Mezi alternativy patří také nástroje založené na platformě Eclipse[6], jako je Modelio, Papyrus nebo OpenCABE. OpenCABE je předchozí nástroj vyvinutý v rámci naší výzkumné skupiny, se kterým se však kvůli omezením a komplexnosti platformy Eclipse ani po 6 letech vývoje nepodařilo dosáhnout stavu, kdy by studenti mohli nástroj použít pro vlastní projekty vyžadující implementaci nových modely i funkce. To se s OpenPonk podařilo již v několika případech.

OpenPonk je založen na myšlence jednoduchého rozšiřitelného jádra, tedy základních tříd a podpory konceptuálního modelování a dále rozšiřitelného pomocí pluginů zajišťujících rozšíření o nové modely, notace a algoritmy dodatečně vytvořené uživateli (uživatel se zpravidla myslí uživatel-vývojář, který pro svoji práci vyvíjí vlastní pluginy OpenPonk).

2 Architektura

OpenPonk slouží primárně pro vývoj nástrojů pracujících nad modely s grafickou reprezentací – diagramy. Pro úpravy modelu pomocí diagramů je jádro tvořeno dle principů model-view-controller (MVC)[12].

Model je v MVC základním kamenem obsahujícím doménový model. Modelem je v našem případě meta-model určité diagramové notace -- například UML meta-model diagramu tříd dle specifikace[9]. Pokud je MVC model (tedy meta-model diagramu) vytvářen přímo s ohledem na implementaci v OpenPonk, stačí využít předpřipravené třídy a pouze doplnit specifiky pro daný model. Zajímavou možností je použít již existující model, který neobsahuje základní funkce potřebné pro použití v OpenPonk. V takovém případě musí za tyto funkce převzít odpovědnost controller (viz dále), k čemuž je možné využít technologii MetaLinks[4]. To dovolí plnou integraci modelu bez nutnosti jeho úprav, což se podařilo u FAMIX[5] modelu pro UML diagramy tříd (viz dále).

View zajišťuje vyobrazení modelu prostřednictvím jednotlivých elementů na vykreslovací ploše. Ty poté reprezentují zpravidla konkrétní prvek modelu. K tomu využijeme grafickou knihovnu pro práci s vektorovou grafikou Roassal[1]. Umožňuje snadnou tvorbu nových tvarů, interakcí s nimi a dalších úprav. U jednodušších notací, jako jsou konečné automaty, postačují tato rozšíření Roassalu, avšak komplexnější notace vyžadují implementovat další vrstvu nad Roassalem.

Tvorba a správa vizuálních entit se provádí v *controlleru*, kde je po uživateli požadována implementace několika metod za pomoci metod a tříd připravených v jádře OpenPonk a v knihovně Roassal. Zpravidla má samotný model i každý jednotlivý prvek modelu svůj vlastní controller. Controllery jsou zodpovědné za interpretaci signálů uživatele přicházejících z view a propagaci změn z a do modelu. Controller se tedy stará o propojení view pro daný model, ale také o ovládání formuláře pro úpravu údajů a palety nástrojů pro ovládání a tvorbu prvků diagramu. V OpenPonk nemá view přímý

přístup k modelu, protože o jejich interakci se plně stará právě controller. To je vhodné zejména pro využití již existujících modelů, které nebyly na podobnou integraci navrženy.

Zodpovědností controllerů je také průběžná validace propojení prvků pomocí hran a vložení prvků do jiných již během jejich tvorby, kdy pro základní funkce stačí implementovat několik málo metod požadovaných třídami jádra OpenPonk. Pro komplexní validace může uživatel vyvinout vhodné rozšíření, jako je například OntoUML validation editor[18].

Uživatelské rozhraní je implementováno ve frameworku *Spec*[13]. Základní *Spec* okno obsahuje ovládací prvky aplikace a *Spec* podokno *Editor*, který zahrnuje vykreslovací plochu, paletu nástrojů pro práci s diagramem, formulář pro úpravy údajů a další podokna vázaná k diagramu. Každé *Spec* podokno obsahuje API pro jeho úpravy.

3 Možnosti rozšíření a využití

Každé rozšíření – plugin – je podtřídou předpřipravené třídy *Plugin*. Pro zahájení práce s modelem, controllery a vizuálními prvky stačí implementovat základní funkčnost, poté lze realizovat i různá rozšíření funkčnosti OpenPonk pro daný typ modelu. Dále uvádíme několik příkladů.

3.1 Úprava modelu pomocí skriptů

Kromě klasického přístupu ke tvorbě instancí modelů (modelů konkrétních diagramů) pomocí vizuálních nástrojů mají současné nástroje také možnost tvorby a úprav pomocí skriptů v jazycích vyvinutých speciálně pro tento jediný účel. Příkladem je *Epsilon Object Language*[7] na platformě *Eclipse*. Vývojář však musí tento skriptovací jazyk vytvořit a spravovat a uživatel se ho musí učit a potýkat se s limity takového jazyka. OpenPonk je však vyvinuta v živém prostředí *Pharo*, což umožňuje manipulaci přímo pomocí programovacího jazyka *Pharo Smalltalk*, který má uživatel k dispozici.

3.2 Vizuální simulace

Při očekávaném rozšíření funkčnosti je nutné vytvořit odpovídající API, které má však omezené možnosti a vyžaduje dodatečnou správu. Např. realizovaný simulátor konečných automatů používá přímý přístup k modelu a pohledu, které o tomto rozšíření nemusí vědět. Simulátor získává informace z modelu a na základě toho upravuje vizuální vrstvu (pohled).

3.3 UML round-trip engineering pro platformu ABM Cormas

Ve spolupráci s výzkumnou skupinou *Cirad RU Green* jsme vyvinuli editor UML diagramů tříd pro platformu *ABM Cormas*[3][17], obsahující podporu round-trip engineering, tedy možností tvořit kód na základě modelu a naopak. Cílem není zcela automatický převod, ale poskytnutí pomoci s tvorbou struktury kódu nebo diagramu. Právě zde

je využito propojení s FAMIX modelem, aniž by tento model bylo nutné přizpůsobovat pro použití s OpenPonk.

Podpora pro zpětnou tvorbu modelu z vygenerovaného kódu vyžaduje některé informace, které v kódu nemusí být obsaženy, což platí o to více v dynamicky typovaných jazycích. Aby kód tyto informace obsahoval, tak musí generátor kódu přidat dodatečné informace pro tyto účely. Rozložení prvků diagramu je pak tvořeno automaticky[16].

4 Závěr

Jelikož naše výzkumná skupina pracuje s různými formami a využitími konceptuálního modelování, je pro nás OpenPonk velmi významným nástrojem.

OpenPonk je úspěšným pokračováním našeho předchozího nástroje na platformě Eclipse, OpenCABE[10], který posloužil jako zdroj nejlepších architektonických principů. Ty byly použity a upraveny tak, aby vznikl vysoce upravitelný a rozšiřitelný nástroj s jednoduchým jádrem. Z názvu OpenPonk je patrná část „otevřený“. Toho je dosaženo i díky otevřené, dynamické a živé platformě Pharo.

Nástroj je v mnoha ohledech stále v rané fázi vývoje, avšak již nyní byl úspěšně použit pro další projekty. Rozšiřitelnost a jednoduchost principů OpenPonk se ukázala i ve studentských projektech, kde se studenti byli schopni úspěšně seznámit s platformou a využít ji pro své práce spočívající v implementaci nové notace či nových algoritmů. Naším cílem je poskytnout komunitě nástroj pro výzkum, vývoj a experimentování, k čemuž se přibližujeme s každým novým uživatelem.

V současnosti spolupracujeme s nizozemskou společností ForMetis Enterprise Engineers na vývoji simulací a validací průmyslových modelů a jsme v kontaktu s INRIA Lille Nord Europe a Univerzitou v Antverpách, které se zajímají o bližší spolupráci.

Další informace o OpenPonk je možné nalézt v obsáhlejší článku v anglickém jazyce[15].

Poděkování: V současné době je vývoj OpenPonk sponzorován firmou ForMetis Consultants. Vývoj podpory round-trip engineering (zejména editoru UML digramu tříd pro ABM CORMAS) byl financován díky RU Green CIRAD. Vývoj MetaLinks byl sponzorován společností Synectique a ESUG prostřednictvím programu Mobility Support.

Literatura

1. Bergel, A.: Agile Visualization. (2016). Available at: agilevisualization.com.
2. Bergel, A., Cassou, D., Ducasse, S., Laval, J., Bergel, J.: Deep into Pharo. Square Bracket, [S.l.], (2013). ISBN 978-3-9523341-6-4.
3. Bommel, P., Becu, N., Le Page, C., Bousquet, F.: Cormas, an Agent-Based simulation platform for coupling human decisions with computerized dynamics. (2015). Available at: <https://agritrop.cirad.fr/576753/2/CormasforIsaga2015.pdf>.
4. Denker, M.: Sub-method Structural and Behavioral Reflection. PhD. thesis, University of Bern, (2008).

5. Ducasse, D., Anquetil, N., Bhatti, M. U., Hora, A. C., Laval, J., Girba, T.: MSE and FAMIX 3.0: an interexchange format and source code model family. (2011). Available at: <https://hal.inria.fr/hal-00646884/>.
6. Eclipse: Graphical Modeling Project. (2016). Available at: <https://eclipse.org/modeling/gmp/>.
7. Kolovos, D., Rose, L., Garcia-Dominguez, A., Paige, R.: The Epsilon Book, volume 20. (2016).
8. MetaCase: MetaEdit+. (2016). Available at: <http://www.metacase.com/>.
9. Object Management Group: OMG Unified Modeling Language 2.5. (2015). Available at: <http://www.omg.org/spec/UML/2.5>.
10. Pergl, R., Tůma, J.: OpenCASE a tool for ontology-centred conceptual modelling. In Advanced Information Systems Engineering Workshops. Springer, (2012), pp. 511–518.
11. Podloucký, M., Pergl, R.: Towards Formal Foundations for BORM ORD Validation and Simulation. SCITEPRESS - Science and Technology Publications, (2014), pp. 315–322. ISBN 978-989-758-027-7 978-989-758-028-4 978-989-758-029-1. doi: 10.5220/0004897603150322.
12. Pope, S. T., Krasner, G. E.: A Cookbook for Using Model-View-Controller User Interface Paradigm in Smalltalk-80. (1988).
13. Van Ryseghem, B., Ducasse, S., Fabry, J.: Seamless composition and reuse of customizable user interfaces with Spec. Science of Computer Programming, (2014), 96:34–51.
14. Sparx Systems: Enterprise Architect. (2016). Available at: <http://www.sparxsystems.com/products/ea/index.html>.
15. Uhnák, P., Pergl, R.: The OpenPonk modeling platform. Proceedings of International Workshop on Smalltalk Technologies (IWST'16), (2016).
16. Uhnák, P.: Layouting of Diagrams in the DynaCASE Tool. Bachelor's thesis, Czech Technical University in Prague, Faculty of Information Technology, (2016).
17. Uhnák, P., Bommel, P.: Facilitating the Design of ABM and the Code Generation to Promote Participatory Modelling. Submitted for publication, (2016).
18. Vološin, M.: Vizualizace instancí OntoUML modelů. Diplomová práce, České vysoké učení technické v Praze, Fakulta informačních technologií, Praha, (2016).

Annotation:

OpenPonk – A Conceptual Modelling Platform for Education, Research and Practice

OpenPonk, formerly known as DynaCASE, is an emerging open software platform for conceptual modelling. Its goal is to support a user-friendly diagram creating and editing and further models validations, transformations and other algorithms. Working with domain-specific languages is also supported. Currently, OpenPonk contains support for finite state machines, Petri net, BORM ORD, DEMO, OntoUML and UML Class Diagrams in various stages of maturity. The vision is to offer an open, easily extensible platform for implementing modelling notations and algorithms. Researchers, teachers and practitioners are the target community.

Compared to other current solutions, which are usually based on Java/Eclipse/EMF/GMF, our solution is implemented in a pure object-oriented technology Pharo/Roassal, which is considerably simpler to master and extend by "watch and learn" and "copy-paste". Moreover, our solution is a live system, where the user may interact with the models during the development.

We also present a project for French institute CIRAD, which is based on OpenPonk.

We recommend paper in the English language[15] for more information.

RDF Storage for Semantic Big Data Historian

Václav Jirkovský^{a,b}, Martin Possolt^a, Marek Obitko^b

^aCzech Institute of Informatics, Robotics and Cybernetics
Czech Technical University in Prague
Žitkova 4, Prague, Czech Republic

^bRockwell Automation Research and Development Center
Pekařská 695/10a, Prague, Czech Republic

{vaclav.jirkovsky, martin.possolt}@cvut.cz
mobitko@ra.rockwell.com

Abstract. Nowadays, data acquisition, subsequent processing as well as analytics are gaining importance within the industrial automation domain. We are witnessing the trend of replacing traditional approaches with more capable Big Data methods and paradigms. To exploit this change, we propose so called Semantic Big Data Historian to handle data of larger volume, variety and velocity. Our historian software prototype benefits from ontological data model and from Hadoop platform. In this paper, we describe briefly our implementation of a prototype of such historian software. Next, we introduce possible models of storage layer of our proposed Semantic Big Data Historian with respect to RDF data format. The storage models exploit Distributed File System of the corresponding Big Data framework for storing RDF data. Finally, the proposed and implemented hybrid model is introduced together with the possible extensions.

Contribution type: Research paper

Key words: Ontology, RDF, Big Data, Hadoop

1 Introduction

The trend of exploiting Big Data paradigms and technologies is coming to the domain of industrial automation for sensor data acquisition, processing and acquisition. This approach for data management and processing offers usable ways how to solve obstacles such as huge data volumes, requirements for “real-time” data processing, and increasing data heterogeneity. Furthermore, the requirements for processing and analyzing heterogeneous data sources are more evident in this domain, and therefore data have to be accompanied by their semantic description including their mutual relations.

To address these needs, we have proposed and implemented Semantic Big Data Historian (SBDH) [1] which is able to handle semantics and Big Data as well. The core data we are storing are data from sensors. For describing both the sensors and data we

use the SHS (Semantics for Historian Sensors) ontology which is based on SSN (Semantic Sensor Network) ontology¹. We have extended existing concepts and relations to capture the data that we need to process.

One of the most important SBDH issues is a realization of the SBDH storage layer which is responsible for RDF data handling. As we already indicated, the solution exploits a Big Data framework for data processing – specifically Apache Hadoop². Thus, we discuss possibilities how to store RDF data by means Hadoop Distributed File System in this paper. The data nature (predominantly time-series) is taken into consideration.

2 HDFS Model

Hadoop Distributed File System (HDFS) is scalable and reliable data storage designed to run on commodity hardware [2]. HDFS is similar to other distributed file systems, but the key differences are as follows – highly fault-tolerant, designed for low-cost hardware, provides high throughput access to application data, and relaxes a few POSIX³ requirements to enable streaming access to file system data. Next, HDFS is able to manage large files and therefore is suitable for storage layer of the Semantic Big Data Historian.

We identified three different approaches of possible HDFS utilization for storing RDF data according to the way they handle data model:

- **Single file model:** preserves the triple construct of classical RDF.
- **Vertical partitioning model:** splits RDF triples according their property.
- **Entity class-based model:** utilizes high-level entity class graph to create RDF partitions [3]. First, similar entities (subjects) are grouped (according to similarity measure) into an entity class. Corresponding entity class graph is then partitioned using METIS⁴. This model is not discussed in the following sections because it is not used in SBDH.

2.1 Single File Model

The *single file model* preserves the RDF triples in the form (*subject, predicate, object*). In other words, data are stored within HDFS in one file. The HDFS is then responsible for splitting the file into blocks, replicating the blocks, etc.

The system based on this approach and HDFS is for example PigSPARQL [3]. Furthermore, SHARD [4] uses a variation of the single file model where triples with the same subject are merged into a one line of a HDFS file.

¹ <https://www.w3.org/2005/Incubator/ssn/ssnx/ssn>

² <http://hadoop.apache.org>

³ POSIX – The Portable Operating System Interface

⁴ <http://glaros.dtc.umn.edu/gkhome/metis/metis/overview>

```
:CO2ds048 rdf:type :CO2ObsValue :hasQuantityValue "355.0"
           :hasQuantityUnitOfMeasurement :parts-per-million
```

2.2 Vertical Partitioning Model

The previously introduced single file model is easy to implement but has some disadvantages. The main obstacle is the I/O cost during query processing. A more suitable model is represented by *vertical partitioning model*. In this model, triples are partitioned with the respect to their property and stored in files named according the corresponding property name. The vertical partitioning model is employed for example in [5]. In the case of SBDH, the file *hasQuantityUnitOfMeasurement* contains the following data:

```
:CO2ds048 :parts-per-million
:THSds075 :percentage
:THSds075 :degreeCelsius
:PRSds032 :hectopascal
```

This model overcome deficiencies of the single file model but data are not homogeneously distributed in files in some cases (e.g., the *type* file is usually very big file). In the case of SBDH, the biggest file would be *hasQuantityValue*.

Further file splitting can be performed for ensuring homogeneous data distribution among files. HadoopRDF⁵ creates partitions according to data property and object as well. For example, the triple (*:THSds075 :hasQuantityUnitOfMeasurement :percentage*) would be stored in a file named *hasQuantityUnitOfMeasurement#percentage*.

3 Hybrid SBDH Model

The current realization of the SBDH storage architecture is based on combining single file model and vertical partitioning-like model. This hybrid model replaced previously used single file model which was insufficient due to the high I/O costs during query processing. The single file model is unsuitable for time-series data storage. Especially for queries with range filter expressions and order constraints.

In detail, the vertical partitioning is used for all sensors measurements where the partitions are created with the respect to *subject* and *property* accompanied by timestamp. For example, the file *CO2ds048#hasQuantityValue* contains the following data:

```
2012-04-29T00:00:10 355.0
2012-04-29T00:00:40 355.1
2012-04-29T00:01:10 355.0
```

⁵ <http://cs.utdallas.edu/semanticweb/Hadoop-RDF/hadoop-rdf.html>

Other triples are stored according to the single file model. The different data handling of sensors measurements reflects the fact that an amount of measurements is significantly bigger than the rest of data.

4 Conclusion

In this paper, we have tackled the problem of the RDF data storage for Semantic Big Data Historian. We shortly described possible ways how to store data by means of HDFS. Then, we briefly introduced the hybrid model of SBDH storage layer.

The combination of semantic description of industrial data together with exploitation of the Hadoop framework represents important step towards a scalable, robust, distributed, and fault tolerant data processing and analytical solution. Such a solution is essential for allowing more efficient and more useful decision making.

Based on our preliminary measurements, we conclude that the hybrid model is satisfactory for SBDH storage layer. However, there are still some deficiencies. As our future work, we are planning to optimize time-series storage. Samples querying could be sped up by utilization of more advanced data structure where data are grouped by particular time interval – e.g., measurements can be grouped in hierarchical structure by year, month, day, hour, etc.

Acknowledgement: This research has been supported by Rockwell Automation Laboratory for Distributed Intelligent Control (RA-DIC) and by institutional resources for research by the Czech Technical University in Prague, Czech Republic.

References

1. M. Obitko and V. Jirkovský, "Big Data Semantics in Industry 4.0," in *HoloMAS 2015*, Springer International Publishing, 2015, pp. 217-229.
2. D. Borthakur, "HDFS architecture guide," HADOOP APACHE PROJECT http://hadoop.apache.org/common/docs/current/hdfs_design.pdf, p. 39, 2008.
3. A. Schätzle, M. Przyjaciół-Zablocki and G. Lausen, "PigSPARQL: Mapping SPARQL to pig latin," in *Proceedings of the International Workshop on Semantic Web Information Management*, 2011.
4. K. Rohloff and R. E. Schantz, "High-performance, massively scalable distributed systems using the MapReduce software framework: the SHARD triple-store," in *Programming Support Innovations for Emerging Distributed Applications*, ACM, 2010, p. 4.
5. M. Husain, J. McGlothlin, M. M. Masud, L. Khan and B. M. Thuraisingham, "Heuristics-based query processing for large RDF graphs using cloud computing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 9, pp. 1312-1327, 2011.
6. X. Zhang, L. Chen, Y. Tong and M. Wang, "EAGRE: Towards scalable I/O efficient SPARQL query evaluation on the cloud," in *ICDE*, 2013.

Modelování a transformace fiskálních datasetů technologiemi RDF v projektu OpenBudgets.eu

Jakub Klímek^{1,2,3}, Jindřich Mynarz¹, Vojtěch Svátek¹

¹FIS VŠE v Praze, Nám. W. Churchilla 4, Praha 3, ČR

²FIT ČVUT, Thákurova 9, Praha 6, ČR

³MFF UK Praha, Malostranské nám. 25, Praha 1, ČR

`jakub.klimek@fit.cvut.cz {jindrich.mynarz,svatek}@vse.cz`

Abstrakt. Příspěvek se věnuje dvěma aspektům tvorby propojitelných dat z oblasti rozpočtů a výdajů veřejné správy (fiskálních dat), která je součástí náplně projektu Horizont 2020 OpenBudgets.eu. Jedná se o návrh datového modelu, postavený na aplikaci RDF slovníku Data Cube Vocabulary (DCV) a kódovnicích ve formátu SKOS, a o převod dat (typicky ve formátu CSV) do cílového formátu RDF pomocí transformačních aplikací, zejména tzv. ETL frameworku LinkedPipes. Součástí převodu je i validace dat systémem integritních omezení.

Typ příspěvku: Aplikační příspěvek

Klíčová slova: RDF, rozpočet, Data Cube Vocabulary, propojování

1 Úvod

Data o rozpočtech a výdajích veřejné správy (fiskální data), jsou z technického pohledu soubory pozorování popsaná hodnotami určitých charakteristik (dimenzí). Typicky se jedná o dimenze časové, prostorové (geografické), administrativní, tematické, apod. Data lze reprezentovat jako vícerozměrné "kostky" a následně analyzovat prostředky datové analytiky (interaktivní vizualizace, data mining apod.).

V projektu Horizont 2020 OpenBudgets.eu (2015-2017) je cílem podpořit různé scénáře využívání fiskálních dat ze strany novinářů, nevládních organizací bojujících proti korupci (např. Transparency International) i místních občanských aktivistů – specificky v kontextu tzv. participativní tvorby rozpočtu. Tyto tři oblasti proto byly zvoleny jako modelové cílové úlohy, s odpovídajícími pracovními balíčky v *aplikační* části projektu. Jednotlivé pracovní balíčky *technologické* části projektu se pak postupně věnují 1) definici datového modelu, 2) extrakci, předzpracování a propojování dat, a automatizovaným analytickým úlohám nad nimi, 3) vizualizaci dat a výsledků analýz, a 4) integraci vyvinutých nástrojů do jednotné platformy. V tomto příspěvku se věnujeme prvnímu a částečně druhému okruhu aktivit, na kterých se zásadním způsobem podílí tým z ČR (pod hlavičkou VŠE Praha, ale s faktickým zapojením expertů i z MFF UK a FIT ČVUT, tj. jde o společné úsilí akademických partnerů iniciativy *OpenData.cz*).

Potenciál fiskálních dat pro analytické úlohy se výrazně zvyšuje jejich obohacením o data podrobněji popisující objekty, kterých se týkají jednotlivé dimenze datové kostky. Geografické lokality (obce, regiony) tak mohou kupříkladu být zahrnuty spolu se svými demografickými, ekonomickými (např. HDP na osobu) a politickými (např. dominantní politická strana) charakteristikami. Vzhledem k potenciálu takového propojování byly v projektu pro modelování a zpracování dat zvoleny technologie založené na jazyku RDF.

2 Datový model fiskálních dat

Klíčovou roli při modelování dat sehrává RDF slovník Data Cube Vocabulary (DCV) [1] a velké počty kódovníků veřejné správy převedených do formátu SKOS. Základní datový model OpenBudgets.eu zahrnuje celkem 20 komponent odpovídající specifikaci DCV:

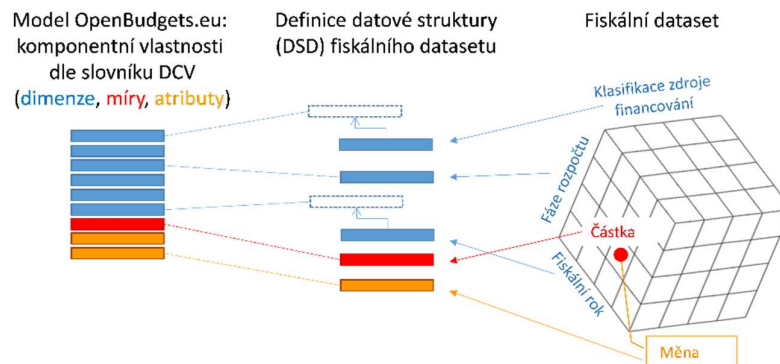
- 17 *dimenzí*, definujících zejména fiskální období, administrativní, ekonomickou nebo funkční klasifikaci výdaje, fázi přípravy/realizace rozpočtu, příjemce platby, asociovaný projekt, nebo datum vzniku výdaje. Hodnoty dimenzí jsou zpravidla vybírány z odpovídajících kódovníků.
- 2 *atributy* (měnu platby a odlišení, zda platba zahrnuje zdanění)
- 1 *míru* (samotnou finanční částku).

Standardní postup využití datového modelu typicky zahrnuje následující kroky:

1. Analýzu vstupního datového souboru (typicky v tabulkovém formátu CSV nebo na něj přímočaře převoditelném): významu jeho jednotlivých sloupců a v nich používaných hodnot.
2. Určení, které ze sloupců je vhodné zařadit do cílového formátu dat, a jakému typu komponenty (dimenze, atribut, míra) v daném kontextu odpovídá.
3. Přímé namapování těch komponent z modelu OpenBudgets.eu, které dostatečně odpovídají významu sloupců.
4. Vytvoření nových komponent, a to pokud možno jako specializací (podvlastností) stávajících komponent, a jejich namapování na sloupce.
5. Sestavení tzv. definice datové struktury (DSD) z namapovaných komponent.
6. Převod dat do formátu RDF, ve struktuře odpovídající vytvořené DSD.

Vztah mezi množinou komponent, DSD a fiskálním datasetem o třech dimenzích je schematicky naznačen na Obr. 1. V prostřední části schématu nahoře je naznačeno odvození nové vlastnosti vyjadřující „klasifikace zdroje financování“, z obecnější komponenty modelu OpenBudgets.eu, vyjadřující „ekonomickou klasifikaci“.

Použitelnost datového modelu byla ověřena jak v rámci projektu jako takového, tak i v prostředí výuky (předmětu zaměřeného na otevřená propojená data – linked data). Na základě tohoto ověření byla zformulována soustava integritních omezení, odchyťující časté modelovací chyby. Jako takové chyby byly identifikovány zejména:



Obr. 1: Vztah modelu OpenBudgets.eu, DSD a fiskálního datasetu

- přímé využití abstraktních komponent (např. „klasifikace“) které mají být pouze specializovány novými komponentami (např. „ekonomická klasifikace“)
- chybějící povinná komponenta (např. „fiskální období“ nebo „měna“)
- vytvoření nové komponenty ve jmenném prostoru základního modelu (tzv. „namespace hijacking“)
- vytvoření komponenty podřazením nesprávnému typu z DCV
- použití vlastního kódovníku pro dimenzi, která má již kódovník definován (v takovém případě má být zavedena nová dimenze jako podvlastnost).

3 Automatická transformace dat

Zpracování probíhá v transformačních aplikacích, zejména v tzv. frameworku LinkedPipes ETL¹ [2]. Data jsou extrahována z primárních zdrojů (tabulky v CSV, případně struktury v XML) a vyjádřena pomocí obecných komponent DCV doplněných o prvky specifické pro fiskální data, včetně kódovníků jednotlivých dimenzí, jsou následně validována a propojována na externí data. Zpracování je popsáno pomocí strukturovaných procesů (pipelines), které mohou být opakovaně aktivovány pro podobné vstupní datasety, průběžně monitorovány a modifikovány.

Jako součást procesů automatické transformace byla implementována i výše zmíněná integritní omezení vyjádřená v dotazovacím jazyce SPARQL. Aktuálně řešenou úlohou je pak využití tzv. linkovacích nástrojů pro automatické vytváření propojení datových prvků (zejména kódovníkových hodnot dimenzí) na externí datové sady. LinkedPipes ETL má takové nástroje, např. Silk,² již začleněné do inventáře použitelných prvků, v rámci tzv. Data Processing Units.

¹ <http://etl.linkedpipes.com/>

² <http://silkframework.org/>

Struktura datového modelu i proces získávání a využití zpětné vazby jsou podrobněji popsány v příspěvku na konferenci SEMANTiCS 2016 [3]. Aktuální stav podpory tvorby pipelines v prostředí LinkedPipes ETL, na základě tzv. „pipeline fragments“, včetně specifických pro model OpenBudgets.eu, pak lze nalézt v příspěvku přijatém na workshop SemStats 2016 [4].

4 Závěr

Projekt OpenBudgets.eu usiluje o systematickou podporu tvorby propojitelných dat v oblasti veřejných rozpočtů a výdajů. Na základě úvodní fáze prací, stručně shrnuté v tomto příspěvku, jsou v současnosti uskutečňovány ukázkové automatické analýzy a vizualizace, a ty z nich, které budou mít širší využitelnost, budou následně integrovány do dlouhodobě udržované softwarové platformy.

Poděkování: Tato publikace vznikla s částečnou podporou projektu EU Horizont 2020 č. 645833 (OpenBudgets.eu).

Literatura

1. Cyganiak, R., Reynolds, D.: The RDF Data Cube Vocabulary. W3C recommendation, W3C, 2014.
2. Klímek, J., Škoda, P., Nečaský, M.: LinkedPipes ETL: Evolved linked data preparation. In The Semantic Web: ESWC 2016 Satellite Events - ESWC 2016 Satellite Events, Anissaras, Crete, Greece, May 29-June 2, 2016, Revised Selected Papers, to appear 2016.
3. Mynarz, J., Svátek, V., Karampatakis, S., Klímek, J., Bratsas, C.: Modeling fiscal data with the Data Cube Vocabulary. In: SEMANTiCS 2016, Posters and Demos, CEUR-WS, to appear 2016.
4. Mynarz, J., Klímek, J., Dudáš, M., Škoda, P., Engels, C., Musyaffa, F. A., Svátek, V.: Reusable transformations of Data Cube Vocabulary datasets from the fiscal domain. In: 4th Int'l Workshop on Semantic Statistics (SemStats 2016), collocated with ISWC 2016. CEUR WS, to appear 2016.

Annotation:

Modeling and transforming fiscal datasets using RDF technology in the OpenBudgets.eu project

The paper addresses the topic of constructing linked datasets in the fiscal domain, as addressed by the (Horizon 2020) OpenBudgets.eu project. The two aspects concerned are: 1) the design of a data model based on the RDF-based Data Cube Vocabulary (DCV) and on code lists in the SKOS format, and 2) the transformation of source data (typically in the CSV format) to the target RDF format using transformation applications (esp. the ETL framework LinkedPipes ETL). The transformed data is also validated by a system of integrity constraints.

Grafová databáze jako úložiště metadat pro data lineage – zkušenosti a výzvy

Karel Quast, Michal Valenta

Fakulta informačních technologií
České vysoké učení technické v Praze
Thákurova 9, 160 00 Praha 6, Česká republika

{karel.quast, michal.valenta}@fit.cvut.cz

Abstrakt. V projektu zabývajícím se tzv. data lineage jsme již před 3 lety pro datové úložiště metadat použili grafovou databázi namísto relační. Od té doby přibýlo instalací, zvětšil se objem dat a změnil se požadavky zákazníků. Řešili jsme temporální dimenzi úložiště a také více možných pohledů na hierarchii dat, tedy kromě fyzické hierarchie, která vychází z analýzy příslušných datových slovníků, přidat například hierarchii logickou/konceptuální. Grafové databáze se pro tento typ úlohy ukazují jako perspektivní řešení. V příspěvku přiblížíme požadavky na datové úložiště metadat pro data lineage, podělíme se o vlastní zkušenosti s praktickou implementací a naznačíme další směry rozvoje a výzvy, které s nimi souvisí.

Typ příspěvku: Příspěvek o probíhajícím výzkumu

Klíčová slova: grafová databáze, temporální databáze, data lineage

1 Úvod

Z hlediska aplikace (praktického použití) souvisí naše práce s dynamicky se rozvíjející oblastí řízení dat (Data Governance) v oblasti datových skladů (Data Warehouses). Konkrétně se zabýváme návrhem a implementací metadatového úložiště pro sledování tzv. „datové linie“ (data lineage).

Pojem data lineage v Encyclopedia of Database Systems [1] odkazuje k pojmu „původ dat“ (Data Provenance). Ten je zaveden následovně: „The term “data provenance” refers to a record trail that accounts for the origin of a piece of data (in a database, document or repository) together with an explanation of how and why it got to the present place.“ V naší realizaci metadatového úložiště jsme skutečně schopni sledovat původ dat – tedy odkud (z jakého zdroje) konkrétní položka pochází, budeme se však raději držet pojmu data lineage, protože je v oblasti datových skladů a nástrojů nad jejich metadaty běžnější a je intuitivně lépe srozumitelný.

Formální zavedení pojmu data lineage společně se základní dvoustupňovou kategorizací lze najít v technickém reportu [2]. Kontext pro zavedení pojmu data lineage je

definován následovně: máme množinu vstupních dat I_1, \dots, I_k , která vstupují do (obecného) grafu transformací T_1, \dots, T_n , ze kterého vychází množina výstupních dat O_1, \dots, O_m . V tomto kontextu pak zkoumáme a popisujeme individuální transformace – data lineage.

Zpráva dále nabízí dvě základní hlediska pro kategorizaci problému data lineage. Prvním hlediskem je převažující způsob dotazování - „*where*“ ptáme se hlavně na cestu, kterou data v systému prochází nebo „*how*“ kdy dáváme větší důraz na porozumění tomu, jak se data mění v průběhu zpracování. Druhé hledisko zavádí kategorie „*schema*“ nebo „*instance*“ podle toho co sledujeme. Z této klasifikace pak vychází mnohé další práce – například [3,4].

Příspěvek je dále členěn takto: v kapitole 2 přiblížíme prostředí, ve kterém úložiště realizujeme, kapitola 3 je věnována zkušenostem, které s realizací dosud máme, poslední, čtvrtá kapitola, nastiňuje další směr výzkumu a experimentů.

2 Projekt Manta a požadavky na metadatové úložiště

Metadatové úložiště, o kterém pojednává tento příspěvek, je jádrem skupiny nástrojů Manta Tools¹. Jedná se o produkt společnosti Profinit určený k vizualizaci data lineage, řízení dat (data governance) a analýze SQL kódu v (heterogenním) prostředí produkčních databází a datových skladů větších podniků (banky, telekomunikační operátoři apod.), na jehož rozvoji se ČVUT FIT podílí v rámci projektu TAČR.

Z pohledu klasifikace data lineage problémů uvedeného v předchozí kapitole, se ve většině případů užití nástrojů Manta pohybujeme v kvadrantu daném kategoriemi „*where*“ - akcentujeme tedy zobrazení linie původu dat spíše než popis toho, jak se data mění a „*schema*“ - zajímají nás prvky (databázové) struktury, nikoliv konkrétní hodnoty (instance) – proto metadatové úložiště.

Z rodiny nástrojů Manta Tools se dále soustředíme na nástroj Manta Flow. Ten pracuje v následujících krocích:

1. Analýza datových slovníků všech databází v organizaci. Výsledkem jsou hierarchie objektů – například technologie – databáze – schéma – tabulka – sloupec, technologie – databáze – schéma – package – procedura – parametr apod. Jednotlivé stromy spojíme jedním zastřešujícím uzlem, čímž vznikne jedna hierarchie (strom) která odráží fyzickou strukturu uložení dat v organizaci.
2. Analýza ETL procesů a transformačních skriptů. Každý takový skript přidá orientované hrany mezi listy stromu. Každá hrana znázorňuje, že příslušný element se mění na jiný (sloupec tabulky se stává sloupcem pohledu, sloupec pohledu se stává vstupním parametrem procedury, sloupec textového (CSV) souboru se stává sloupcem tabulky apod.).

¹ <https://mantatools.com/products/>

3. Vzniklý (obecný) graf popisuje jak fyzické uložení dat v databázích, tak datové toky v organizaci. Tento graf je základem pro vizualizace datových toků na různých úrovních (sloupců, objektů schématu, databází,...) a dalších analýz datových toků (dopadové analýzy, bezpečnostní audit apod.).

Na počátku projektu byl pro úložiště metadat použit relační databázový stroj PostgreSQL. Pro (aktuálně) středně velké nasazení nástroje (řádově cca 1 milión uzlů a 3 milióny hran) byla odezva databázového stroje ještě dostačující, očekávalo se však navýšení množství dat.

Výše popsaná struktura úložiště metadat – obecný orientovaný graf – je totožná s datovým modelem tzv. grafových databází – důležitým členem pestré rodiny tzv. NoSQL databází, které se v posledních 10 letech velmi bouřlivě rozvíjejí, hledají a často i nacházejí své uplatnění ve specifických aplikačních doménách. Nejen struktura, ale i typické dotazování – hledání cesty v grafu nebo sledování datového toku od nějakého konkrétního elementu dále přímo vybízí k nasazení tohoto typu databázového stroje.

3 Nasazení grafové databáze a další rozvoj úložiště

V této kapitole nejprve v sekci 3.1 velmi stručně popíšeme zkušenosti s nasazením grafových databází, dále pak dvě úpravy datového úložiště – v části 3.2 to bude přidání temporální dimenze, která umožní sledovat vývoj datových toků v čase a v části 3.3 přidání dalších pohledů na hierarchii datových objektů.

3.1 Nasazení grafové databáze

Experimentovat jsme začali s databázovým strojem Neo4j. Použili jsme testovací databázi s cca 1 miliónem uzlů a 3 milióny hran. Můžeme konstatovat, že prostorové nároky na uložení dat byly u Neo4j cca o 25% větší než u PostgreSQL, s rostoucí velikostí databáze se zvětšovaly zhruba stejně. Podobné to bylo s importem dat (PostgreSQL byl v průměru o cca 10% rychlejší).

V souladu s očekáváním zvítězila grafová databáze u složitějších dotazů (širší okolí uzlu s případnou následnou filtrací). Zřejmě není ani překvapivé, že použití specializovaného dotazovacího jazyka Cypher, který nabízí velmi pohodlné dotazování, bylo o řád pomalejší než použití přímého Java API. Z hlediska požadavků na úložiště i rychlost zpracování se databáze Neo4j ukázala jako vhodná. Detailní popisy měření a výsledky lze najít v [7].

Další experimenty proběhly s grafovým DB strojem Titan. Oproti Neo4j má velmi zajímavý rys – podobně jako v případě MySQL je možné si zvolit datové úložiště (v době experimentování byly k dispozici binární úložiště PersistIt, BerkeleyDB a Cassandra).

Při testování vykázal Titan podobné výsledky jako Neo4j a nakonec byl v projektu nasazen s použitím binárního úložiště PersistIt.

3.2 Přidání časové dimenze

Další optimalizace zvoleného úložiště se zaměřila na použití konkrétních indexů vhodných pro naše využití. Tyto experimenty a také přidání časové dimenze jsou podrobně popsány v diplomové práci [5]. Zde se omezíme na konstatování, že dále používáme podpůrné (externí) indexování Elasticsearch a informace o časové platnosti objektů jsou udržované na hranách grafu.

3.3 Přidání dalších hierarchií

Analýza, návrh a implementace více pohledů - hierarchií dat včetně rozsáhlého měření je popsána v diplomové práci [6]. Zde se pouze omezíme na konstatování, že jsme zvolili variantu, kdy odlišná hierarchie je vyjádřena pomocí hrany.

4 Další rozvoj úložiště

Do budoucna plánujeme vytvoření komplexnějšího benchmarku pro testování tohoto typu úložiště a dále zkoumat možnosti efektivní distribuce úložiště (požadavky větších zákazníků si to vyžadují) a paralelizaci importů analyzovaných transformačních skriptů.

Literatura

1. Ling Liu, M. Tamer Özsu (editors): Encyclopedia of Database Systems. ISBN: 978-0-387-35544-3 (Print) 978-0-387-39940-9 (Online).
2. Robert Ikeda and Jennifer Widom. Data lineage: A survey. Technical report, Stanford University, 2009.
3. Robert Ikeda, Hyunjung Park, and Jennifer Widom. Provenance for generalized map and reduce workflows. In Proc. of CIDR, January 2011.
4. Dionysios Logothetis, Soumyarupa De, and Kenneth Yocum. 2013. Scalable lineage capture for debugging DISC analytics. In Proceedings of the 4th annual Symposium on Cloud Computing (SOCC '13). ACM, New York, NY, USA, Article 17, 15 pages.
5. Petr Holeček: Temporální data v grafové databázi v projektu Manta. Diplomová práce. České vysoké učení technické v Praze, Fakulta informačních technologií, Praha, 2015.
6. Michal Peroutka: Optimální struktura a indexy modelu metadatového úložiště v grafové databázi. Diplomová práce. České vysoké učení technické v Praze, Fakulta informačních technologií, Praha, 2016.
7. Michal Valenta: Návrh datového úložiště projektu Nástroje pro automatizaci Quality Assurance rozsáhlých Business Intelligence systémů a datových skladů. První výroční zpráva projektu TAČR. 2014.

Annotation:

Graph database as a storage for data lineage metadata – experiences and challenges

We used a graph database as a storage for data lineage metadata instead of relational one in a project 3 years ago. New product installations appeared, amount of data increased, and user requirements changed during this time. We had to design our own solution for temporal dimension of the storage and we are working on multiple hierarchy data view model (i.e. allow for example a logical/conceptual hierarchy alongside with implicitly used physical one. It seems graph databases are suitable DBMS for this kind of usage.

We are presenting some implementation specific details about our approach and share some experiences related to particular SW used in the project. We also try to scratch challenges we are facing now and corresponding ways of research.

Životní situace jako základní výchozí bod eGovernmentu

Václav Řepa

Katedra informačních technologií fakulty informatiky a statistiky
Vysoká škola ekonomická v Praze
nám.W.Churchilla 4, 130 67, Praha 3, Česká republika

repa@vse.cz

Abstrakt. Životní události (životní situace) jsou obvykle chápány jako zvláštní pohled na činnosti veřejné správy, blízký jejím klientům: občanům. Tento pohled obvykle pomáhá efektivně a pro klienty srozumitelně organizovat webové platformy veřejných institucí. Avšak skutečný význam životních událostí je mnohem podstatnější. Jde o pohled, kdy činnosti veřejné správy vnímáme jako důsledky skutečných událostí v reálném životě klientů veřejné správy. Příspěvek představí přístup k analýze životních situací v rámci modelování životních cyklů objektů veřejné správy a využití tohoto přístupu v reálném probíhající projektu. Budou též diskutovány důležité souvislosti, jako je vztah životních událostí k procesům ve veřejné správě, jakož i jejich vztah k e-governmentu, a to včetně ilustrací na příkladech ze zmíněného projektu.

Typ příspěvku: Výzkumní příspěvek

Klíčová slova: Životní situace, veřejná správa, konceptuální modelování, procesní řízení.

1 Úvod

Veřejnou správou (VS) rozumíme správu věcí veřejných, jež *objektivně* vyplývá z potřeby péče o hodnoty, které přesahují rozměr individua. Tato potřeba plyne z faktu, že podstata člověka je společenská, fungující společnost je základním předpokladem, prostředím a také místem osobní realizace každého člověka. Nelze přitom spoléhat na automatickou shodu osobních zájmů a konání se zájmy společnosti jako celku. Na druhou stranu existují individuální potřeby, problémy a situace, které daná osoba není schopna řešit vlastní silou, přestože mohou být pro ni kriticky důležité, až fatální. To vše určuje potřebu a smysl existence veřejné správy.

Společné hodnoty, tvořící podstatu potřeby VS, dělíme do tří oblastí: *Fyzickým prostředím* rozumíme dané území a jeho přírodní, a další fyzické hodnoty, o něž je třeba pečovat. *Sociální prostředí* představuje lidi a jejich osobní a společenské, tedy kulturní a další hodnoty, důležité pro fungování společnosti. *Podnikatelským prostředím* pak rozumíme veškeré hodnoty, potřebné pro využívání příležitostí k reali-

zaci hodnot. V těchto třech oblastech je objektivní potřeba veřejného konání. Základním úkolem VS je konat ve smyslu péče o hodnoty ve výše uvedených třech základních dimenzích života společnosti, jež se vzájemně prolínají, ovlivňují a podmiňují. V tomto pojetí tedy VS není něčím daným, definovaným a konservovaným jednou provždy legislativou, nařízeními, či dokonce technologií, ale dynamickým systémem péče o dynamicky se rozvíjející svět věcí veřejných, kde každý jeden příslušník společnosti má svou danou osobní odpovědnost a s tím i související práva. *Schopnost dynamiky v jedné VS je hlavním determinantem dynamiky rozvoje celé společnosti.* Aby byla VS dostatečně dynamická, přizpůsobivá měnícím se podmínkám¹, musí být postavena na základních obsahových prvcích života samotného, které existují relativně nezávisle na změnách prostředí a jsou základním obsahovým vymezením smyslu činnosti VS. Za tyto základní prvky považujeme klíčové události v životech jednotlivých aktérů a objektů společnosti – tzv. **životní situace**.

2 Konceptuální analýza oblasti působení veřejné správy jako základ analýzy životních situací

Pojem *životní situace* (nebo také *životní události*) je ve veřejné správě používán zhruba od konce devadesátých let. Původní, a stále v podstatě jediný používaný, význam tohoto pojmu souvisí s tvorbou webových portálů organizací veřejné správy. Životní situace představují původně netradiční pohled na veřejnosprávní „agendy“ - pohled z pozice potřeb / problémů jejího klienta. V jediném, již neexistujícím, britském projektu LEAP [6] byly životní situace pojímány nám podobným způsobem, tedy jako základní prvky životního cyklu objektu VS (s tím, že zde šlo o jediný objekt *Občan*). Náš přístup je tak ve světovém měřítku v podstatě jedinečný. V projektu obskurního názvu „Optimalizace životních situací ve vztahu k registru práv a povinností“ [2], vedeném na Ministerstvu vnitra ČR v roce 2015 v rámci rozvojových projektů EU, se neočekávaně podařilo prosadit myšlenku, že nezbytným základem ke koncepci VS, zejména v duchu tzv. eGovernmentu, kdy maximum rutinních akcí VS má být převzato technologií, je **objektivní představa životních situací**, jejich základní seznam a vědomí jejich důležitých souvislostí. Taková objektivní představa musí vzejít z dostatečně exaktní analýzy původu takových situací. Proto se stala základem zmíněného projektu **konceptuální analýza prostředí, v němž má VS konat**. Jelikož fakticky jde o obecný model reálného života ve výše zmiňovaných oblastech (viz Obrázek 1), nazýváme jej: **„Ontologický model oblasti působení veřejné správy“**.

Model je veřejně pozorovatelný na webu [1] a je vyveden v jazyku UML [5] s inspirací z jeho rozšíření OntoUML [3], pro modelování složitých vztahů objektů (viz níže). Používá Class Diagram pro systémový / globální pohled na reálné objekty a dále State Chart pro detailní pohledy na životy jednotlivých klíčových objektů. Volba UML je motivována jednak obecně, faktem, že se jedná o oborový informatický modelovací

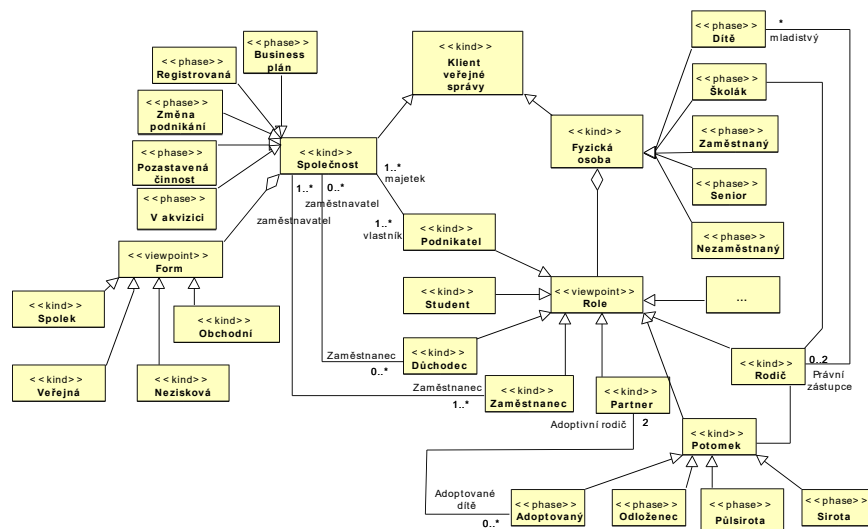
¹ A to včetně nových technologií, jež vytvářejí především možnosti konat zcela jinak, než tradičním způsobem.

standard², jednak specificky objektovým charakterem UML, umožňujícím nejen existenci, ale i vývojový (životocykelný) pohled na objekty, jenž se právě v ontologii oblasti zájmu veřejné správy ukázal být klíčově důležitým.

Zatímco výše zmiňované 3 základní oblasti působení VS jsou externími, objektivně danými zdroji životních situací, životní cyklus objektu těmito situacím dává *subjektivní kontext příslušného individua*. A právě zohledňování individuálních hodnot je jedním z kritických problémů (potažmo výzev) současné veřejné správy. Význam jedné a téže objektivní události je tak pro různé objekty velmi rozdílný, a to právě proto, že je v rozdílných kontextech jejich individuálních životů. Jedna objektivní událost, životní změna (např. *Dosažení plnoletosti*) tak typicky znamená množství životních situací rozdílných významů pro různé dotčené objekty (*Rodič, Potomek, Škola,...*). Smyslem modelování životních cyklů konceptuálních objektů je tak vyjádřit obecné zákonitosti kontextu jejich životních situací a tím obecně postihnout individuální různost významů téže fyzické události pro různé její aktéry.

2.1 Modelování životních situací

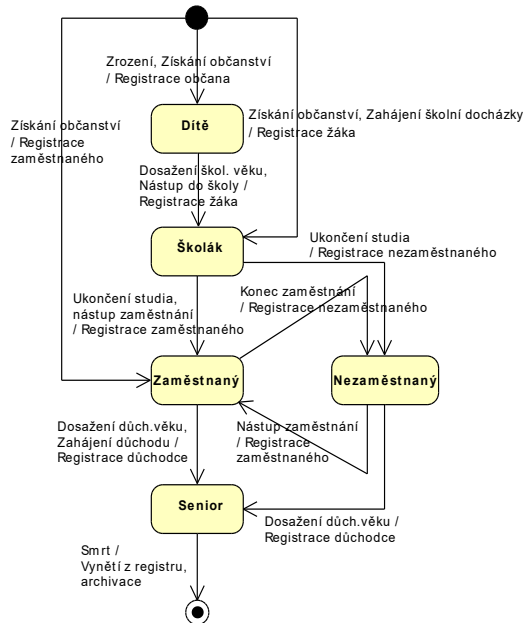
Vedle *class diagramu* (viz Obr. 1) pro systémový pohled na objekty a jejich vztahy byl k modelování životního cyklu objektu, coby soustavy životních situací, použit další klíčový diagram jazyka UML: *state chart* (viz Obr. 2).



Obr. 1 Systémový model objektů zájmu veřejné správy – fragment

2 I vzhledem k tomu, že projekt je součástí směřování k tzv. eGovernmentu, tedy by měl být schopen být základem k implementaci aplikací IT ve VS, resp. být s nimi integrován.

Samotné životní situace pak byly modelovány jako základní prvky popisu životního cyklu – přechody mezi jeho jednotlivými stavy. Podle definice stavového diagramu UML (*state chart*) je každý přechod mezi stavy popsán dvojicí údajů: událost (externí podnět k přechodu mezi stavy) a akce (tomu odpovídající metoda ze životního cyklu objektu). První uvedený údaj je životní událostí, zatímco ten druhý představuje vazbu na příslušné reakce veřejné správy na danou životní událost / situaci (vazbu na potřebné procesy VS). Zde uvedené příklady jsou zjednodušeným výsekem z původního modelu, pozorovatelného na [1]. Z modelu na Obr. 2 je vidět, jak jsou zachyceny základní, obecně platné, časové a kauzální zákonitosti životních událostí, vázaných k jednomu objektu. Jednotlivé popsané přechody mezi stavy vymezují nutné / toliko možné vzájemné časové kombinace životních událostí. Například je vidět, že dosažením školního věku již daná osoba nikdy nebude mít šanci být dítětem, nebo že ze stavu Nezaměstnaný lze uniknout pouze buď získáním zaměstnání, nebo dosažením důchodového věku, v němž, byť důchodce může být zaměstnán, ztráta zaměstnání, jakkoliv reálně může nastat, již není, z hlediska veřejné správy, relevantní životní událostí, vyžadující reakci, apod. Na Obr. 2 je také mj. vidět, že nezávislé fatální události (zde Smrt), nemohou být modelovány v rámci životního cyklu, jsouce zcela nezávislými na ostatních událostech (prakticky tu chybí, na základě této události reálně možné, přechody ze všech stavů do stavu terminálního). Zohlednění této události vyžaduje ještě vyšší abstrakci, trivializující celý život osoby do jediného stavu *Živá*, jehož jsou všechny, zde uvedené stavy, součástí. V našem modelu je tato informace obsažena v (rovněž triviálním) životním cyklu Klienta veřejné správy.



Obr. 2 Životní cyklus objektu *Fyzická osoba*

Obrázky 1 a 2 také ukazují způsob použití obou diagramů ve vzájemné souvislosti. Každý přechod mezi stavy objektu vždy odpovídá nějakému vztahu k jinému objektu (asociaci, nebo příslušnosti ke generalizační, či agregační struktuře). S tím souvisí významný metodický přínos tohoto způsobu modelování ontologie: *poznávání životních cyklů, vzájemných kauzálních a časových závislostí jednotlivých událostí jeho životního cyklu, je mocným nástrojem rozvoje poznání nutných vztahů mezi objekty modelu.*

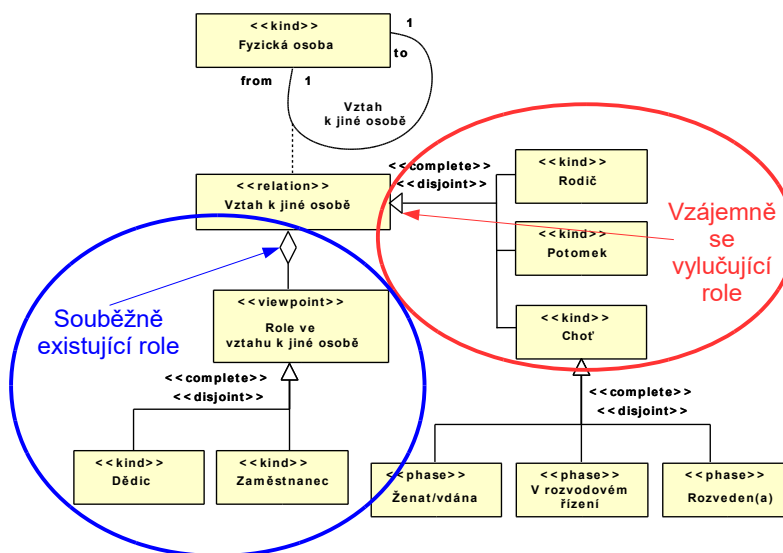
3 Závěr: metodické důsledky analýzy životních situací

Nehledě na značný význam projektu pro společnost a pojetí veřejné správy, zejména pak pro záměr tzv. eGovernmentu, jež nejsou primárním předmětem zájmu tohoto příspěvku, měl projekt značný vliv na metodický rozvoj v oblasti konceptuálního a ontologického modelování. Ukázalo se především, že pro potřebu modelování životních situací ve veřejnosprávním významu, je třeba vidět v reálném světě jen několik málo fyzických objektů, považovatelných za základní (na Obr. 1 jsou to v podstatě jen dva základní druhy Klienta VS – Společnost a Fyzická osoba). Drtivá většina objektů, jejichž životní cykly je třeba modelovat, jsou abstraktní objekty, představující různé významy objektů základních (jejich role) v různých vzájemných souvislostech a z různých úhlů pohledu (viz všechny ostatní objekty na Obr. 1). Bylo přitom nutno překonat rozpor mezi generalizačním a agregačním významem téže struktury, představující jednak specializaci významů a současně i jednotlivé fáze, agregované v životním cyklu téhož objektu. Je přitom nutno překonat rozpor mezi generalizačním a agregačním významem téže struktury, představující jednak specializaci významů a současně i jednotlivé fáze, agregované v životním cyklu téhož objektu. K tomu účelu byla vytvořena řada strukturních vzorů, postavených na základních stereotypech objektů: Kind-Phase, Kind-Kind, Kind-Viewpoint-Phase a Relation-Viewpoint-Kind-Phase, jež jsou popsány v metodické dokumentaci k nalezení na [2]. Vzory volně vycházejí z jazyka OntoUML [3] a rozšiřují jej o problematiku modelování dynamiky (časových aspektů) objektů. Zmíněné strukturní vzory pak také ukazují možný směr vhodného rozšíření jeho meta-ontologie UFO [4].

Obrázek 3 ukazuje příklad použití vzoru Relation-Viewpoint-Kind-Phase. Jde o nejsložitější ze zmíněných vzorů, z nich složený a v modelu nejčastěji potřebný. Používá se pro modelování vztahů mezi objekty (stereotyp <<relation>>). Mezi těmiž dvěma objekty typicky existuje souběžně množství různých vztahů a současně až několik skupin vztahů, jež se vzájemně vylučují. Každý jeden vztah může být zapotřebí (a patrně, až na malé výjimky, bude) modelovat životním cyklem. Každý vztah k jinému objektu zpravidla představuje jistou roli, kterou modelovaný objekt, ve vztahu k tomu druhému, hraje. Objekt je tedy modelován souhrnem svých paralelních životů v různých hraných rolích. Jednotlivé souběžné životy jsou na sobě buď nezávislé – jsou vzájemně asynchronní (ty jsou modelovány jako agregát různých úhlů pohledu - <<viewpoint>>), nebo se vzájemně vylučují (ty jsou pak modelovány jako prostá specializace stereotypů <<kind>>). Příklad na Obr. 3 ukazuje současně nezávislé role Dědice a Zaměstnanec (lze zaměstnat svého dědice) a vylučující se role Rodiče Potomka a Chotě

(rodič nemůže být dítětem svého potomka, ani s ním nesmí vstoupit v manželství). Případné dílčí závislosti životů (synchronizaci) pak modelujeme obecnou asociací mezi danými objekty (pokud by například dědictví bylo podmíněno manželstvím, resp. jeho specifickým průběhem apod.).

V budoucnu plánujeme, kromě pokračování na obsahu modelu, jej validovat meta-ontologií UFO [4] a zahájit tím mezinárodní spolupráci na jeho dalším rozvoji.



Obr. 3 Použití vzoru Relation-Viewpoint-Kind-Phase

Literatura

1. Ontologický model VS: http://ontologie_vs.panrepa.org/.
2. Projekt OŽSvVkrPP: <http://www.mvcr.cz/clanek/optimalizace-zivotnich-situaci-ve-vztahu-k-registru-prav-a-povinnosti.aspx>.
3. Guizzardi, G.: 'Ontological Foundations for Structural Conceptual Models', Telematica' Institut, Fundamental Research Series No. 15, ISBN 90-75176-81-3 ISSN 1388-1795, 2005.
4. Guizzardi, G., Wagner, G., Almeida, J.P.A., Guizzardi, R.S.S: 'Towards ontological foundations for conceptual modeling: The unified foundational ontology (UFO) story' in Journal of Applied Ontology, vol. 10, no. 3-4, pp. 259-271, 2015, DOI: 10.3233/AO-150157.
5. Object Management Group: UML 2.0 Superstructure Specification, Doc.# ptc/03-08-02, Aug. 2003.
6. The Local Authority EMAS and Procurement (LEAP) Project (<http://www.leap.gov.uk/>)

Annotation:

Life events as a basic starting point of eGovernment

Life events (life situations) are usually understood as a specific view on public administration activities which is close to their clients: citizens. This view usually helps to effectively and client - friendly organize the web platform of the public authority. Nevertheless, the real meaning of life events is more essential. Such a view allows regarding the public administration activities as consequences of real events in real lives of public administration clients. The paper introduces the approach to the analysis of life situations in the context of life cycles of the public administration objects and the use of this approach in the real ongoing project. The relation of life events to the public administration processes as well as their relation to the e-Government are discussed and illustrated with examples from the project.

Explorace společných charakteristik ontologií formální konceptuální analýzou

Ondřej Zamazal, Vojtěch Svátek

Fakulta informatiky a statistiky
Vysoká škola ekonomická v Praze
nám. W. Churchilla 1938/4, 130 67 Praha 3, Česká republika

{ondrej.zamazal, svatek}@vse.cz

Abstrakt. Znalostní ontologie jsou na webu dostupné v různých kolekcích. Kolekce bývají využívány k výběru ontologií pro testování sémanticko-webových nástrojů. Některé kolekce umožňují vyhledávání ontologií pomocí atributů jakými jsou například počty tříd a fulltextové vyhledávání podle klíčových slov. Výběr jednotlivých ontologií je tak umožněn do té míry do jaké atributy odrážejí požadavky na hledané ontologie. Vedle toho existují práce zaměřující se na agregované statistiky kolekcí ontologií s cílem umožnit rozlišení kolekcí ontologií. Zatímco vyhledávání ontologií specifikováním hodnot atributů naráží na omezující nutnost jasné představy nastavení těchto atributů, vyhledávání ontologií pomocí agregujících popisných statistik naráží na omezující zobecňování ontologií v kolekci. V tomto příspěvku předkládáme metodu explorace společných charakteristik ontologií formální konceptuální analýzou. Formální konceptuální analýza uspořádává množiny objektů z hlediska jejich společných charakteristik do podoby formálních konceptů v konceptuálním svazu. Průzkum konceptuálního svazu tak může podpořit výběr ontologií z kolekcí tím, že ukazuje souvislosti mezi různými kombinacemi atributů a odpovídajícími objekty.

Typ příspěvku: Příspěvek o probíhajícím výzkumu

Klíčové slova: ontologie, sémantický web, formální konceptuální analýza

1 Úvod

Na jednu stranu na sémantickém webu přibývají znalostní ontologie a na druhou stranu se neustále objevují nové nástroje, které tyto ontologie využívají. Nové nástroje přirozeně potřebují testovat svoji funkčnost na rozličných ontologiích. Za tímto účelem vznikají různé možnosti nalezení vhodných ontologií.

Znalostní ontologie lze najít v klasických kolekcích ontologií nebo pomocí vyhledávačů. Mezi nejznámější vyhledávače ontologií patří Watson,¹ který shromažďuje ontologie a další sémantické dokumenty z webu a umožňuje vyhledávání pomocí klíčových

¹ <http://watson.kmi.open.ac.uk/>

slov z různých aspektů ontologií, např. popisky (labels). Watson také nabízí programové rozhraní, pomocí něhož lze získat hodnoty některých metrik např. počty prvků ontologií. Podle těchto metrik, ale nelze ontologie vyhledávat.

Vedle vyhledávačů sbírajících ontologie volně z webu jsou vytvářeny kolekce ontologií, které se soustředí na kvalitní ontologie z jedné oblasti, např. BioPortal² z oblasti biomedicíny a nebo na kvalitní ontologie specificky používané, např. ontologie z Linked Open Vocabularies (LOV)³ používané pro popis propojených dat na webu. Obě kolekce nabízejí možnost vyhledávání podle klíčových slov a získání základních charakteristik. Ani tyto kolekce však nemají možnost, jak vyhledávat podle charakteristik.

Dále jsou vytvářeny nástroje, které jednak počítají souhrnné statistiky ontologií z původních kolekcí a jednak je zpřístupňují pro další použití. Matentzoglou et al. v [2] představil nástroj⁴ pro sdílení a tvorbu kolekcí ontologií. Nástroj obsahuje sourné statistiky a zprostředkovává kolekce ontologií BioPortal, Oxford Ontology Library, Tones a vlastní kolekci ontologií MOWLCorp. Sestavení vlastní kolekce ontologií je omezeno na vyplnění "offline" HTML formuláře s několika parametry. Možnost "online" vyhledávat ontologie a sestavovat z nich testovací kolekce podle mnoha (kolem 70) charakteristik nabízí nástroj „Online Ontology Set Picker“ (OOSP)⁵ [4].

Zatímco vyhledávání ontologií specifikováním hodnot atributů prakticky naráží na omezující nutnost jasné představy nastavení těchto atributů, vyhledávání ontologií pomocí agregujících popisných statistik naráží na omezující zobecňování ontologií v kolekci.

V tomto příspěvku se zabýváme explorační společných charakteristik ontologií formální konceptuální analýzou jako další možností usnadnění výběru ontologií z různých kolekcí. Formální konceptuální analýza uspořádává množiny prvků z hlediska jejich společných charakteristik do podoby formálních konceptů v konceptuálním svazu. Průzkum konceptuálního svazu tak může podpořit výběr ontologií z kolekcí tím, že ukazuje souvislosti mezi různými kombinacemi atributů a odpovídajícími objekty (ontologiemi). Tyto souvislosti tak mohou být vzaty v potaz při nastavování atributů při vyhledávání kýžených ontologií v situaci, kdy se cílí na ontologie bohatě zastoupené různými atributy při jejich dostatečném množství.

2 Explorace ontologií formální konceptuální analýzou

Formální konceptuální analýza (FKA) [1] představuje jednu z metod explorativní analýzy tabulkových dat. Výhodou FKA je získání nových netriviálních poznatků o vstupních datech. Výstupem je tzv. konceptuální svaz jako hierarchicky uspořádaná množina shluků neboli formálních konceptů z dat na vstupu. V základní podobě FKA pracuje s objekty, které mají bivalentní logické atributy (ano/ne atributy). Podle (ne)přítomnosti atributů jsou objekty rozděleny do formálních konceptů (pojmů), které lze chápat jako

² <http://bioportal.bioontology.org/>

³ <http://lov.okfn.org/>

⁴ <http://mowlrepo.cs.manchester.ac.uk/>

⁵ <http://owl.vse.cz:8080/OOSP/>

dvojice (A, B) , kde A je množina objektů a B je množina atributů, které patří pod pojem. Dále musí platit, že A jsou objekty, které všechny mají atributy z B a současně B je množina atributů společná všem objektům z A . Objekt-atributová data pouze s ano/ne hodnotami představují základní formální kontext. V případě, že potřebujeme vícehodnotové atributy jedná se o vícehodnotový kontext, který je pomocí konceptuálního škálování převeden na základní kontext pro použití FKA.

V našem případě ontologie představují objekty a atributy odpovídají metrikám ontologií. Celkem pracujeme se šesti skupinami ontologických metrik (např. metriky týkající se entit). Všechny atributy jsou vícehodnotové a pro aplikaci základní FKA je nutné nejprve převést vícehodnotový kontext na základní konceptuálním škálováním. V případě numerických atributů pro tvorbu škály používáme diskretizační metodu ekvifrekvenčních intervalů. Všechny atributy jsou následně převedeny na bivalentní varianty. Pro generování konceptuálních svazů používáme specifikaci minimální podpory v datech jednotlivými koncepty.

Exploraci jsme provedli nad ontologiemi z LOV kolekce (509 ontologií) dostupné přes nástroj OOSP. Na základě testování jsme došli k nastavení maximálního počtu ekvifrekvenčních intervalů na 5 a minimální podpory 50 ontologií. Na jednu stranu při vyšším počtu ekvifrekvenčních intervalů byla nedostatečná podpora v datech a výsledný konceptuální svaz měl plochou strukturu. Na druhou stranu nastavení nižší podpory by znamenalo příliš malou extenzi nalezených konceptů. Při vybraném nastavení maximálně 5 intervalů bylo vytvořeno 167 bivalentních atributů a výsledný konceptuální svaz obsahoval 5 úrovní. Příkladem nalezeného konceptu na třetí úrovni je např. $\{labels [7,20), range class [2,7), object property range [2,7)\}$ (51), který zahrnuje 51 ontologií s relativně malým počtem popisků (labels), pojmenovaných tříd v oborech hodnot objektových vlastností (range class) a objektových vlastností s definovaným oborem hodnot (object property range). Uvedené intervaly zahrnují nižší hodnoty příslušných metrik, kdežto většina konceptů v konceptuálním svazu obsahuje spíše intervaly s extrémními hodnotami jako např. koncept ze čtvrté úrovně svazu: $\{labels [104, 16878], range class [26, 2329], axiom [802, 44101], object property range [27,2329]\}$ (55). V tomto případě koncept obsahuje stejné typy metrik (navíc ještě počty axiomů) ale s nejvyššími hodnotami v intervalech atributů. Tento koncept ukazuje na další typický rys nalezených konceptů při exploraci. Příslušné metriky v nalezených konceptech spolu často úzce souvisejí.

Dominantní přítomnost konceptů s nejvyššími hodnotami intervalů může být částečně způsobena tím, že ontologie s vysokými hodnotami určitých metrik mají také pravděpodobně vysoké hodnoty metrik souvisejících (jako např. počty axiomů a počty objektových vlastností s definovaným oborem hodnot). Dalším ovlivňujícím faktorem je možná míra zkreslení metodou diskretizace na ekvifrekvenční intervaly. V rámci našeho testování ekvifrekvenční intervaly přinášely v procesu explorace smysluplnější výsledky než ekvidistantní intervaly, protože většina metrik má log-normální rozdělení s odlehlými hodnotami. V případě ekvidistantních intervalů většina ontologií byla zahrnuta do několika málo intervalů s nižšími hodnotami a ty pak dominovaly při generování konceptuálního svazu. V případě využití ekvifrekvenčních intervalů pravý krajní

interval zahrnuje "dlouhý chvost" (long tail), což zvyšuje šanci, že více ontologií z daného konceptu jsou spolu v "dlouhém chvostu" i v dalších metrikách a tím se přispívá k dominanci nejvyšších hodnot intervalů v konceptuálním svazu.

3 Závěr

Úvodní explorace pomocí FKA ukázala slibné možnosti nacházení ontologií odpovídajících společným charakteristikám. Další pozornost si zejména zaslouží experimentování s metodou diskretizace, která zásadně ovlivňuje charakter intenzí konceptů. Výsledné koncepty představují kategorizaci ontologií, kterou plánujeme porovnat s výstupy shlukové analýzy [3]. Cílem práce je v budoucnu umožnit uživateli využívat výstupů z FKA nebo jiné data miningové metody jako podpory při sestavování testovací kolekce ontologií.

Poděkování: Ondřej Zamazal byl podpořen z grantu GAČR 14-14076P.

Literatura

1. Bělohávek R.: Konceptuální svazy a formální konceptuální analýza. In: V. Snášel (Ed.): *Znalosti 2004*, 66-84. ISBN 80-248-0456-5.
2. Matentzoglou N., Tang D., Parsia B., Sattler U. The manchester OWL repository: system description. In: Poster a demo sekce ISWC 2014.
3. Rice M. D., Siff M. Clusters, concepts, and pseudometrics. In: *Electronic Notes in Theoretical Computer Science* 40. 323-346. 2001.
4. Zamazal O., Svátek V.: OOSP: Ontological Benchmarks Made on the Fly. In: SumPre Workshop při ESWC 2015. Portoroz, Slovenia. 2015.

Annotation:

Exploration of common characteristics of ontologies by Formal Concept Analysis

Designers of new semantic web tools search for ontologies within different ontology repositories. Repositories differ not only in characteristics of ontologies but also in means how a user can search for suitable ontologies. Some repositories provide an access by specifying values of metrics other enable to use a fulltext search using keywords from various aspects. Other works aim at overall statistics of repositories. While searching for ontologies by a specification of metrics values is restricted due to the fact that a user does not often have an idea of metrics values, using overall statistics of repositories is restrictive due to its generality. This paper deals with an exploratory method of common characteristics for ontologies by Formal Concept Analysis. Formal Concept Analysis organizes a set of objects according to their common characteristics into the concepts within the lattice. An exploration of the lattice might support a selection of ontologies from repositories.

Sociálny web a jeho aplikácie

Šírenie správ a vzťahy medzi užívateľmi v sociálnych sieťach

Eubomír Antoni, Stanislav Krajčí, Ondrej Kridlo

Ústav informatiky, Prírodovedecká fakulta
Univerzita Pavla Jozefa Šafárika v Košiciach
Jesenná 5, 040 01 Košice, Slovenská republika

{lubomir.antoni, stanislav.krajci,
ondrej.kridlo}@upjs.sk

Abstrakt. Sociálnu sieť definujeme ako multirelačnú množinu údajov, ktorú je možné reprezentovať vo forme grafu. V príspevku sa zaoberáme metódami, ktoré umožňujú skúmať vzťahy medzi užívateľmi v sociálnych sieťach. Porovnáваме výsledky, ktoré sme získali longitudinálnou analýzou sociálnej siete žiakov troma rozličnými metódami. V závere formulujeme problém vyhľadávania prekrývajúcich sa komunit z hľadiska šírenia sa správ v sieti.

Typ príspevku: Príspevok o prebiehajúcom výskume

Kľúčové slová: zhľukovanie, prekrývajúce sa komunity, kaskády, faktorizácia matíc

1 Úvod

Z pohľadu dolovania údajov môžeme sociálnu sieť charakterizovať ako heterogénnu a multirelačnú sadu údajov, ktorá je reprezentovaná grafom. Uzly grafu predstavujú objekty, hrany grafu reprezentujú vzťahy medzi tými objektmi alebo interakcie medzi nimi. Sociálne siete môžeme uvažovať nielen v sociálnom kontexte, ale existuje veľa inštancií sociálnych sietí vo svete v podobe technologických, obchodných, ekonomických alebo biologických sociálnych sietí [3].

Komunitu v rámci sociálnej siete zvyčajne definujeme ako skupinu uzlov, ktoré sú husto spojené s ohľadom na zvyšnú časť siete. V prípade, že každý uzol prislúcha v danej sieti len jednej komunite, hovoríme o *disjunktných komunitách*. Mnohé reálne siete sú však charakterizované tým, že uzly siete sú členmi viac než jednej komunity, teda hovoríme o *prekrývajúcich sa komunitách* [4].

V tomto príspevku sa zaoberáme metódami vyhľadávania prekrývajúcich sa komunit. V prvej časti uvažujeme longitudinálnu analýzu žiakov školskej triedy, pričom prekrývajúce komunity sú tvorené žiakmi, ktorí sú podobní z hľadiska ich vzťahov k iným spolužiakom. V druhej časti analyzujeme problém prekrývajúcich sa komunit z hľadiska šírenia sa správ v sieti.

2 Analýza vzťahov medzi užívateľmi

V spolupráci s košickým gymnáziom sme analyzovali zhľady žiakov školskej triedy, ktorí sú si v istom zmysle blízki. Uzly ohodnoteného a orientovaného grafu predstavujú žiakov a hrany sú charakterizované celočíselnými hodnotami v rozsahu od -3 do 3, pričom reprezentujú vzťah hodnotiaceho študenta k spolužiakom [6]. Pri tejto analýze bol použitý jednostranný fuzzy prístup vo formálnej konceptovej analýze [5], ktorý umožňuje generovať prekrývajúce sa komunity. Na redukciu počtu prekrývajúcich komunit sa využíva modifikovaný Rice-Siff algoritmus, ktorý pomocou funkcie vzdialenosti a metrických vlastností umožňuje navyše tieto komunity ohodnotiť z hľadiska ich významnosti. Metóda, ktorá využíva na ohodnotenie významnosti prekrývajúcich sa komunit koncept tzv. horných alfa rezov z teórie fuzzy množín a fuzzy logiky, je prezentovaná v [8]. V práci [1] sme tieto myšlienky modifikovali tak, aby umožňovala výpočty aj v prípade tzv. alfa dolných rezov.

Na základe zozbieraných údajov o žiakoch v rokoch 2007, 2011 a 2014 sme v tomto výskume pokračovali a vyhodnotili navzájom tri vzorky žiakov, nie nutne zhodných. Na analýzu sme použili modifikovaný Rice-Siff algoritmus a metódu horných a dolných alfa rezov, pomocou ktorých môže učiteľ bližšie spoznať štruktúru svojej triedy. Každá z týchto metód zoradí komunity od najvýznamnejšej po najmenej významnú. Pomocou Kendall tau-b koeficientu [9] sme sa snažili odhaliť korelácie, ktoré sú typické pre použité metódy. Oproti tradičnému Spearmanovho korelačného koeficientu, Kendall tau-b koeficient nevyžaduje hodnoty usporiadať podľa veľkosti a priradiť im poradie. Na výpočet Kendall tau-b koeficientu sme použili balíček Kendall v jazyku R, ale vytvorili sme aj vlastnú triedu v Jave na výpočet tohto koeficientu podľa definície, aby sme si potvrdili správnosť výsledku.

Tab 5. Kendallov tau-b koeficienty medzi významnosťou a veľkosťou komunit, resp. medzi významnosťou a obľúbenosťou komunit

| | 2007 | | 2011 | | 2014 | |
|------------|---------|------------|----------|------------|---------|------------|
| | veľkosť | obľúbenosť | veľkosť | obľúbenosť | veľkosť | obľúbenosť |
| Rice-Siff | 0,590** | -0,206 | 0,673** | -0,295** | 0,619** | -0,268 |
| Horné rezy | 0,130** | -0,045 | -0,399** | 0,333** | -0,071* | 0,071* |
| Dolné rezy | 0,181** | 0,196** | 0,306** | 0,203* | 0,252** | 0,181** |

Z tabuľky môžeme vidieť, že významnosť komunit a obľúbenosť žiakov v komunite sú pomocou metódy dolných rezov pozitívne korelované a to signifikantne vo všetkých troch vzorkách. To znamená, že metódou dolných rezov najprv dostaneme žiakov, ktorí sú menej populárni a v ďalších prekrývajúcich sa komunitách už táto popularita narastá. Naopak v Rice-Siff algoritme [5] je silne pozitívne korelovaná významnosť komunit a kardinalita komunit. To znamená, že princíp tohto algoritmu je založený na počiatočnom generovaní malých skupín žiakov, ktoré sú si veľmi blízke a postupne sa kardinalita generovaných skupín zväčšuje.

Neprekrývajúce sa (vzájomne disjunktné) komunity v ohodnotených a orientovaných grafoch je možné vyhľadávať aj napríklad pomocou algoritmu Infomap [7]. Na druhej strane, efektívnu metódu na vyhľadávanie prekrývajúcich sa komunit na základe

vzťahov medzi užívateľmi pre neorientovaný a neohodnotený graf prezentuje [4]. Uvažujeme neorientovaný a neohodnotený graf a jeho maticu príslušnosti, ktorá uchováva informáciu o tom, ktoré uzly sú v grafe navzájom prepojené. Použitím metódy faktori-zácie matíc vieme takúto maticu rozložiť na dve faktorové matice, ktorých súčin najlepšie aproximuje pôvodnú maticu príslušnosti. Vybraný počet faktorov (fixne, resp. na základe vhodnej optimalizácie) zodpovedá počtu nájdených prekrývajúcich sa komunit. Príslušnosť daného uzla v grafe k jednotlivým komunitám určujú hodnoty vygenerované vo faktorovej matici.

3 Analýza šírenia správ medzi užívateľmi

Informácie medzi užívateľmi (virálny marketing) sa šíria ústnou formou, formou odporúčaní na nákup kníh, filmov, ale aj vo forme *informačných kaskád*, napr. na Twitteri. Hovoríme aj o tzv. *sociálnej nákaze*, pri ktorej sa jednotlivci zvyknú prispôbiť správaniu ich rovesníkov [2].

Na reprezentáciu šírenia správ v sociálnej sieti potrebujeme použiť dve grafové štruktúry. Prvou je orientovaný graf, ktorého uzly tvoria užívatelia a orientovaná hrana zodpovedá informácii o tom, že informácia sa šíri od jedného užívateľa k druhému. Druhou štruktúrou je ohodnotený bipartitný graf, ktorý obsahuje dve množiny uzlov (množinu užívateľov U a množinu správ I). Hodnota každej hrany (u, i) medzi užívateľom a správou vyjadruje čas, v ktorom užívateľ u zdieľal správu i svojim nasledovníkom (napr. retweet na Twitteri) [2].

Kaskádou správy i budeme nazývať postupnosť dvojíc (u, t) , kde t vyjadruje čas, v ktorom užívateľ zdieľal správu i . Z množiny kaskád všetkých správ vieme na základe metód popísaných v [10] vybrať len také správy, ktoré spĺňajú istú prahovú hodnotu, napríklad priemernú vzdialenosť medzi nasledovníkmi správy, mieru entropie a podobne.

Keďže medzi kaskádami a komunitami existuje istý vzťah, našim ďalším cieľom je skúmať, akým spôsobom vieme z kaskád identifikovať komunity. Je tiež prirodzené, že hranica danej komunity by mala zastaviť šírenie obvyčajnej správy a jej kaskádu. Schéma na Obr. 1 znázorňuje vzťahy popísané v tomto príspevku, pričom prerušované čiary predstavujú problematiku nášho aktuálneho záujmu a výskumu.



Obr. 1 Schéma analýzy vzťahov a šírenia správ

4 Záver

V práci prezentujeme experiment, v ktorom skúmame prekrývajúce sa komunity žiakov v priebehu niekoľkých rokov. V druhej časti formulujeme úvod do problematiky šírenia správ v sociálnych sieťach, ktorú plánujeme v našom ďalšom výskume doplniť experimentmi a vizualizáciou dát (napr. v systéme Gephi). Detekcia komunit je široká téma, pričom v praxi sú zvyčajne použiteľné metódy s lineárnou komplexnosťou, keďže zložitejšie metódy sú neškálovateľné na reálnych dátach sociálnych sietí.

Podakovanie: Túto prácu podporilo MŠVVaŠ SR v rámci projektu VEGA 1/0073/15 a VEGA 1/0475/14.

Literatúra

1. Antoni, L., Krajčí, S.: Quality measure of fuzzy formal concepts. In: Abstracts of 11th International Conference on Fuzzy Set Theory and Applications, Liptovský Ján, Slovak Republic, (2012), pp. 18.
2. Barbieri, N., Bonchi, F., Manco, G.: Cascade-based Community Detection. In: Proc. ACM Intl. Conf. on Web search and data mining, (2013), pp. 33–42.
3. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Elsevier, (2006).
4. Jin, D., Gabrys, B., Dang, J.: Combined node and link partitions method for finding overlapping communities in complex networks. Scientific Reports, (2015), vol. 5, Article number: 8600.
5. Krajčí, S.: Cluster based efficient generation of fuzzy concepts. Neural Network World, (2003), vol. 13, no. 5, pp. 521–530.
6. Krajčí, S., Krajčiová, J.: Social network and one-sided fuzzy concept lattices. In: Proceedings of FUZZ-IEEE 2007. London, U.K., (2007), pp. 222–227.
7. Rosvall, M., Bergstrom, C. T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences of the United States of America (2008), vol. 105, no. 4, pp. 1118–1123.
8. Snášel, V., Ďuráková, D., Krajčí, S., Vojtáš, P.: Merging concept lattices of alpha-cuts of fuzzy contexts. Contributions to General Algebra, (2004), vol. 14, pp. 155–166.
9. Valz, P.D., Thompson, M.E.: Exact inference for Kendall's S and Spearman's rho. Journal of Computational and Graphical Statistics, (1994), vol. 3, pp. 459–472.
10. Weng, L., Menczer, F., Ahn, Y. Y.: Predicting Successful Memes using Network and Community Structure. In: Proceedings of Eighth International AAAI Conference on Weblogs and Social Media, (2014), pp. 535–544.

Annotation:

Messages spreading and relationships between users in social networks

A social network can be defined as a multirelational data set which is represented by a graph. In this contribution, we present the methods for exploring the relationships between the users of a special social network. We present the comparison of results which we have obtained in the longitudinal study of social networks of students by three different methods. We formulate an issue of finding overlapping communities regarding the information spread in social networks.

Stance detection in online discussions

Peter Krejzl, Barbora Hourová, Josef Steinberger

Department of Computer Science and Engineering, NTIS Center,
Faculty of Applied Sciences,
University of West Bohemia
Univerzitní 8, 306 14, Plzeň
Czech Republic

{krejzl, steinberger}@kiv.zcu.cz
hourova@students.zcu.cz

Abstract. This paper describes our system created to detect stance in online discussions. The goal is to identify whether the author of a comment is in favor of the given target or against. Our approach is based on a maximum entropy classifier, which uses surface-level, sentiment and domain-specific features. The system was originally developed to detect stance in English tweets. We adapted it to process Czech news commentaries.

Contribution type: Work-in-progress paper

Keywords: stance detection, opinion mining

1 Introduction

Stance detection has been defined as automatically detecting whether the author of a piece of text is in favor of the given target or against it. In the third class, there are the cases, in which neither inference is likely. It can be viewed as a subtask of opinion mining and it stands next to the sentiment analysis. The significant difference is that in sentiment analysis, systems determine whether a piece of text is positive, negative, or neutral. However, in stance detection, systems are to determine author's favorability towards a given target and the target even may not be explicitly mentioned in the text. Moreover, the text may express positive opinion about an entity contained in the text, but one can also infer that the author is against the defined target (an entity or a topic). This makes the task more difficult, compared to the sentiment analysis, but it can often bring complementary information [3].

There are many applications which could benefit from the automatic stance detection, including information retrieval, textual entailment, or text summarization, in particular opinion summarization.

2 Task description

2.1 Stance detection at SemEval 2016

The system was originally created for the SemEval 2016 task: Detecting stance in tweets [5]. The task had two independent subtasks – supervised and weakly supervised. The supervised task tested stance detection towards five targets (*Atheism*, *Climate Change is a Real Concern*, *Feminist Movement*, *Hillary Clinton* and *Legalization of Abortion*). Participants were provided 2.814 labeled training tweets for the five targets. In the case of the weakly supervised task, there were no training data but participants could use a large number (around 70K) tweets related to the single target: *Donald Trump*. The goal was to classify tweets into three classes – *IN FAVOR*, *AGAINST*, *NONE*. The performance was measured by the average F1-score on FAVOR and AGAINST classes.

There were 19 participating systems for the supervised subtask and 9 for weakly-supervised subtask. Our system performed well for *Abortion* (2nd), *Climate change* (3rd) and *Hillary Clinton* (4th). The overall rank was 9th. In the weakly-supervised task, we were ranked 4th, only the top system was significantly better. Official results are summarized in the Table 1.

Tab 1. Overall system performance on SemEval’s Twitter data.

| Topic | Our system F1 (rank) | Overall F1 (rank) |
|----------------------------------|----------------------|-------------------|
| Atheism | .5788 (8) | .6342 (9) |
| Climate change is a real concern | .4690 (3) | |
| Feminist movement | .5182 (10) | |
| Hillary Clinton | .5982 (4) | |
| Legalization of abortion | .6198 (2) | |
| Donald Trump | .4202 (4) | .4202 (4) |

2.2 Adaptation to Czech

We used the same system to detect stance in Czech news commentaries. We collected 1.560 comments from a Czech news server¹ related to two topics – “Miloš Zeman” (the Czech president) and “Smoking ban in restaurants” (statistics in Table 2). Consider the following example from the topic “Miloš Zeman”.

Target: *Miloš Zeman*

Comment: „*To je u Zemana běžné, že používá ne pravdy! Viz Peroutka*”². ...³

¹ <http://www.idnes.cz>

² President accused famous journalist Ferdinand Peroutka (1895 - 1978) of supporting Hitler.

³ Can be translated as: “Zeman is doing this normally–using non-truths! For example Peroutka”

Tab 2. Czech news commentaries data – statistics.

| Topic | In favor | Against | None | Total |
|----------------------------|----------|---------|------|-------|
| Miloš Zeman | 180 | 170 | 300 | 750 |
| Smoking ban in restaurants | 170 | 250 | 390 | 810 |

The annotation was done by 2 annotators. There was a fair agreement between them (74%), Kappa was 0.61. The agreement level forms an upper bound for system performance.

3 The approach overview

We preprocessed the Czech commentaries by the same rules as in the original system [3] (for example: all URLs were replaced by keyword ‘URL’, links to images are replaced by ‘IMGURL’, only letters are preserved, the rest of the characters is removed, ...). Moreover, we stemmed the texts by HPS – High Precision Stemmer [2]. The system is based on a standard maximum entropy classifier [4], trained separately for each topic, with the following features.

It has been showed that *unigrams* perform quite well in this task [6]. Our model is based on TF-IDF and uses the top 1000 words from the vocabulary. The rest of the features can be turned on or off for each topic. *Initial n-grams*⁴, as showed in [1] can be useful features. Our system supports initial unigrams to initial trigrams. Another surface feature was the *comment length* in words after preprocessing. We used a resource borrowed from the sentiment analysis – *Entity-centered sentiment dictionaries* (ECSD): dictionaries created mainly for the purpose of entity-related polarity detection [7].

The original system [3] used more features, which could not be easily applied on Czech commentaries. We do not work with tweets, so we could not use a set of features generated from hashtags. We have not analyzed the influence of part-of-speech (POS) tags yet. We did not identify strong candidates to build a domain specific dictionary as in [3]. Bigram features did not work in the case of the tweet analysis, so we did not use it in this work as well. However, we plan to revisit the influence of bigram, POS or domain-specific features.

4 Results

Table 4 shows results on the Czech data. We used two evaluation measures. The first one was used for the SemEval’16 evaluation – the average F1-score on FAVOR and AGAINST classes. The second one includes the NONE class as well. We used 10-fold cross validation to distribute training and testing data.

⁴ Initial n-grams are basically the first n words of the sentence.

Tab 4. System performance on Czech news commentaries.

| Topic | F1 – (In favor/Against) | F1 – (In favor/Against/None) |
|----------------------------|-------------------------|------------------------------|
| Miloš Zeman | .4347 | .5204 |
| Smoking ban in restaurants | .4562 | .5400 |

The results show that performance on the Czech data is significantly worse (.43 – .46) than on the English tweets corpus (.47 – .62). It is mainly due to the lack of some key features like hashtags or domain-specific. Moreover, in the tweets corpus the stance tend to lean to one direction (either FAVOR or AGAINST), while in the Czech corpus most of the comments are considered neutral (NONE).

5 Conclusion

The paper describes the system originally created to participate in Tweet Stance Detection task in SemEval 2016 and additionally used to detect stance in Czech news commentaries. We experienced worse performance in comparison with the original English tweets corpus. It is mainly due to the lack of some significant features like hashtags. The current plan is to revisit the influence of bigram, POS or domain-specific features.

Acknowledgment: This work was supported by grant no. SGS-2013-029 Advanced computing and information systems and by project MediaGist, EU's FP7 People Programme (Marie Curie Actions), no. 630786.

References

1. Anand, P., Walker, M., Abbott, R., Tree, J. E. F., Bowmani, R., and Minor, M. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In Proceedings of WASSA'11, ACL.
2. Brychcín, T., Konopík, M., 2015, HPS: High Precision stemmer. Information Processing & Management, volume 51, Pages 68-91.
3. Krejzl, P., Steinberger, J., 2016, UWB at SemEval-2016 Task 6: Stance Detection, Proceedings of SemEval 2016, pages 408-412, ACL.
4. Loper, E. and Bird, S. 2002. NLTK: The Natural Language Toolkit In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, ACL.
5. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. 2016. Semeval-2016 task 6: Detecting stance in tweets. In Proceedings of SemEval '16, ACL.
6. Somasundaran, S. and Wiebe, J. 2009. Recognizing stances in online debates. In Proceedings of the ACL/AFNLP, pages 226-234, ACL.
7. Steinberger, J., Lenkova, P., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Steinberger, R., Tanev, H., Zavarella, V. and Vazquez, S. 2012. Creating Sentiment Dictionaries via Triangulation. In Decision Support 53(4), pages 689-694, Elsevier.

SoSIREČR – Sociální síť informatiků v regionech České republiky

Jaroslav Pokorný, Peter Vojtáš

Matematicko-fyzikální fakulta
Univerzita Karlova v Praze
Malostranské náměstí 25, Praha, Česká Republika

{pokorny, vojtas}@ksi.mff.cuni.cz

Typ příspěvku: Příspěvek o probíhajícím výzkumu

Klíčová slova: profesionální sociální síť, kvalifikovaný znalostní pracovník, inteligentní vyhledávání pomocí znalostních profilů, Lean startup metodologie, studentské projekty

V tomto příspěvku shrnujeme naše zkušenosti s vývojem a používáním portálu sitit.cz - Sociální síť informatiků v regionech České republiky řešeného v rámci OP VK (Operační program Vzdělávání pro konkurenceschopnost). Po třech letech vývoje je v provozu čtvrtým rokem udržitelnosti.

Zmíníme naše původní příklady použití [6]. Cílem OP VK v ČR je rozvoj vzdělávací společnosti za účelem posílení konkurenceschopnosti ČR prostřednictvím modernizace systémů počátečního, terciárního a dalšího vzdělávání, jejich propojení do komplexního systému celoživotního učení a zlepšení podmínek ve výzkumu a vývoji. Cílem oblasti podpory výzvy (v oblasti podpory 2.4 – Partnerství a sítě) bylo posílení vztahů mezi institucemi terciárního vzdělávání, výzkumnými organizacemi a subjekty soukromého sektoru a veřejné správy. Vzhledem k výzvě a prvotnímu průzkumu zájmu v regionech, byla naše práce orientovaná na podporu spolupráce akademické, podnikatelské a státní sféry s použitím znalostních profilů.

Pokusili jsme se získat další podporu pro rozšíření funkčnosti portálu, ze kterého jsme plánovali vytvořit prostředí pro testování software (typicky on-line uživatelské studie doporučovacích systémů pro webové obchody [5]).

V dalším zmíníme posun v naší orientaci. První se týkal nabídky pro další domény. Implementace našeho portálu je vhodná pro experimentální použití v libovolné znalostně intenzivní doméně (stačí vyměnit XML soubory profilů, [3, 4]). Další se týkal jazykových mutací. Poslední aktivity se týkají podpory imitace studentských start-up vizi podle metodologie Lean startup [7]. Lean startup je metoda pro rozvoj podnikání a produktů poprvé navržena v roce 2008 Ericem Riesem. Na základě své předchozí zkušenosti z práce na jednom softwarovém projektu, Ries tvrdí, že jeho metoda může zkrátit jejich vývojový cyklus výrobku kombinací experimentů založených na byznys-hypotéze, iterativní zveřejňování verzí produktů, a to, co nazývá validovaným učením. Ries tvrdí, že pokud startup investuje čas do iterativního budování produktu resp.

služby na základě požadavků prvních zákazníků (early adopters), může redukovat riziko neúspěchu. V poslední době se objevily i kritické hlasy k Lean startup metodologii, např. [2]. Autoři [2] konstatují, že problém neleží v principech Lean startup-u, ale v jejich použití jako univerzálního receptu na úspěch inovace. Jednoduchá řešení jsou lákavá - ale jsou jen zřídka účinná. Autoři se do této pasti chytili s jejich startupem Gamevy. Jejich zkušenosti jsou hodné zřetele.

Naše experimenty s metodologií Lean startup začaly v rámci výuky předmětů „Sémantizace webu“ a „Uživatelské preference“ [1]. Sociální aspekty portálu jsou použity na imitaci uživatelské zpětné vazby v různých fázích vývoje studentského projektu. Pokrýváme pouze práce od vize po podnikatelský plán (žádné programování). Místo minimálního životaschopného produktu, je úkolem návrh vizuální podoby procesního modelu uživatelského rozhraní. Studenti jsou povzbuzováni k vizím, které řeší klíčová místa problematiky. V sémantizaci webu je to automatizace extrakce a anotace informací z webu. V uživatelských preferencích je to jejich učení z reakcí uživatele.

Poděkování: Portál byl podpořen projektem OP VK č. CZ.1.07/2.4.00/12.0039 a tato práce projektem P-46.

Literatura

1. Pokorný, J., Vojtas, P. Pivoting universal professional social network to help in development of start-up visions, In Dateso 2016
2. Lean Start-Up, and How It Almost Killed Our Company – InfoQ, <https://www.infoq.com/articles/lean-startup-killed>
3. Kubalik, J., Pokorný, J., Vita, M., Vojtas, P. Generic Private Social Network for Knowledge Management. WISE Workshops 2014: 27-41
4. Matousek, K., Kubalik, J., Necaský, M., Vojtas, P. Exploiting Potential of the Professional Social Network Portal "SitIT". ADBIS Workshops 2012: 295-304
5. Kopecký, M., Pokorný, J., Vojtas, P., Kubalik, J., Matousek, K., Maryska, M., Novotný, O., Peska, L. Testing and Evaluating Software in a Social Network Creating Baseline Knowledge. EJC 2012: 127-141
6. Vojtas, P., Pokorný, J., Necaský, M., Skopal, T., Matousek, K., Kubalik, J., Novotný, O., Maryska, M. SoSIREČR - IT professional social network. CASoN 2011: 108-113
7. Ries, E. The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses, Crown Business Publ. 2011

Annotation:

SoSIREČR – a social network of computer scientists in regions of Czech Republic

In this extended abstract we summarize our acquaintance with development and usage of portal sitit.cz - social network of computer scientists in regions of Czech Republic. We describe original use-case and pivoting of our orientation towards support of students' start-up visions.

Twitter a #brexit sentiment

Petr Šaloun, Radka Cepláková

VŠB-Technická univerzita Ostrava,
17. listopadu 15, 708 33 Ostrava, Česká republika

petr.saloun@vsb.cz, r.ceplakova@gmail.com

Abstrakt Sociálne siete v dnešnej dobe hýbu spoločnosťou a sú jedným z hlavných informačných kanálov pre väčšinu populácie. Dianie na sociálnych sieťach je preto zaujímavé sledovať a skúmať. V tweetoch je možné rozpoznať širokú škálu emócií, ktorá sa dá analyzovať a na jej základe určiť aký sentiment prevláda pri učitom hashtagu. Vďaka tomu môžeme vyhodnotiť prípadné bezpečnostné riziko, pre konkrétne osoby alebo miesta. Na testovanie sme si vybrali #Brexit a po dobu troch mesiacov sme zhromažďovali do databázy tweety ktoré obsahovali tento hashtag. Tweety obsahovali, okrem hľadaného, viacero hashtagov a najfrekvencovanejšie príbuzné hashtagy boli #VoteLEAVE, #BrexitNoww a #Euref. Hodnota sentimentu sa pohybovala okolo -830 bodov, a teda je jasné, že používatelia twitteru sú naklonení na jednu stranu. Zistili sme teda, že tweety sú ladené prevažne negatívne, kde negatívita smeruje na EU a tým pádom volia autori k svojmu príspevku minimálne jeden zo spomínaných hashtagov.

Typ príspevku: Work-in-progress paper

Kľúčové slová: tweet, analýza postojov (sentiment), #brexit, natural language processing

1 Úvod

Sociálne siete v dnešnej dobe hýbu spoločnosťou a sú jedným z hlavných informačných kanálov pre väčšinu populácie. Dianie na sociálnych sieťach je preto zaujímavé sledovať a skúmať. V tweetoch je možné rozpoznať širokú škálu emócií, ktorá sa dá analyzovať a na jej základe určiť aký sentiment prevláda pri učitom hashtagu. Vďaka tomu môžeme vyhodnotiť prípadné bezpečnostné riziko, pre konkrétne osoby alebo miesta. V príspevku popisujeme jeden zo spôsobov ako identifikovať bezpečnostné riziko na základe získaných tweetov v anglickom jazyku. V pripravovanej webovej aplikácii bude mať používateľ možnosť do vyhľadávacieho poľa zadať akýkoľvek hashtag a náš systém, na základe týždennej histórie tohoto hashtagu, zanalyzuje tweety aj retweety používateľov Twitteru. Používateľ bude mať možnosť obmedziť rozsah tweetov pre analýzu od niekoľkých dní až do pár hodín. Systém tak môže poskytnúť dlhodobejšiu analýzu, ale aj aktuálny prehľad diania na Twitteri. Po analýze vstupných dát, používateľ dostane prehľadnú štatistiku analyzovaných slov, slovných spojení, príbuzných

hashtagov, percento retweetov a v neposlednom rade sentiment prevládajúci na tomto hashtagu. Keďže pri niektorých hashtagoch nie je zrejmé k akému výsledku sme sa dopracovali ponúkame používateľovi aj náhľad na 5 tweetov, ktoré boli najviac krát retweetnuté. Vďaka príbuzným hashtagom, náhľadu tweetov a bodovému ohodnoteniu bude mať používateľ dobrú predstavu či sú tweety ladené pozitívne, negatívne alebo neutrálne. Nepriamo nadväzujeme na náš predchádzajúci výskum [2, 3].

2 Extrakcia kľúčových slov

Na analýzu tweetov používame metódy spracovania prirodzeného jazyka. Podstatná je extrakcia kľúčových slov, vďaka ktorým môžeme použiť slovníkový prístup a bodovo ohodnotiť analyzovaný tweet. Použitý Nielsenov slovník, pozri [1], obsahuje 2477 slov a ku nim bodové ohodnotenie podľa sily významu, na stupnici od -5 (negatívne) do 5 (pozitívne). Algoritmus vyhľadá slová z tweetu, ktoré sú v slovníku a rozdelí ich na pozitívne ladené alebo negatívne ladené. Na základe tohoto rozdelenia sú potom slovám priradené hodnoty. Súčtom týchto hodnôt a vydelením počtom všetkých slov dostaneme relevantný priemerný sentiment daného tweetu. Proces spracovávania tweetov prebiehal sekvenčne. Na začiatku algoritmus prejde celý tweet a ak obsahuje znaky, ktoré do analýzy nepatria (napr. bodky, čiarky, pomlčky, ...), sú odfiltrované. Tweet je rozdelený na jednotlivé slová, ktoré sú vyhľadané v slovníku a ak slovo slovník obsahuje, sú mu priradené body sentimentu. Ohodnotenie celého tweetu nastáva vtedy, keď všetky slová z tweetu, nájdené v slovníku, majú svoje bodové ohodnotenie. Tieto body sa spočítajú a k tweetu je vypočítané relatívne skóre (skóre vzhľadom ku počtu slov). Vďaka relatívnemu skóre dostaneme relevantný výsledok sentimentu. Po spracovaní tweetu sa uložia text a skóre do databázy.

3 Prípadová štúdia – #Brexit

Na testovanie sme si vybrali #Brexit, pretože ku tomuto hashtagu je denne obrovské množstvo vyjadrení a spoločnosť táto téma zaujíma. Po dobu troch mesiacov sme do databázy zhromažďovali tweety, ktoré obsahovali tento hashtag a na základe získaných dát sme získali celkom zaujímavé výsledky.

Tab 1. Hodnoty sentimentu v priebehu experimentu

| dátum | sentiment hashtagu |
|-------|--------------------|
| 29/04 | -540 |
| 19/05 | -610 |
| 15/06 | -830 |
| 24/06 | 440 |
| 28/06 | 670 |
| 02/07 | 990 |

Tab 2. Počet tweetov

| obdobie | počet tweetov |
|-----------------|---------------|
| pred referendum | 10799 |
| po referende | 10338 |

Tab 3. Počet retweetov

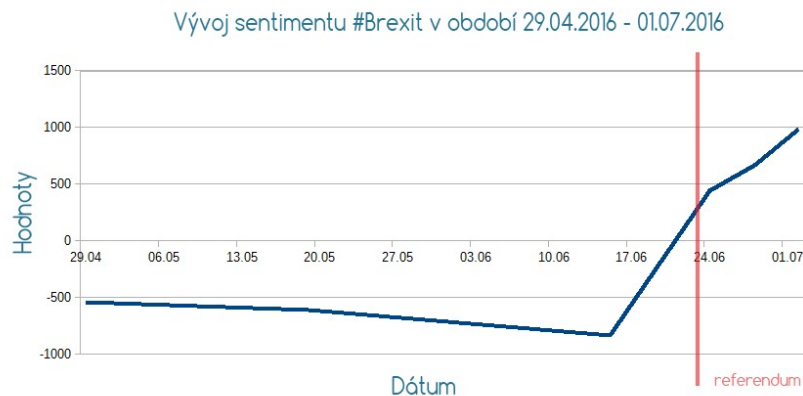
| obdobie | počet retweetov |
|-----------------|-----------------|
| pred referendum | 25.00 % |
| po referende | 83.00 % |

Tab 4. Počet hodnotených slov obsiahnutých v tweetoch

| obdobie | počet hodnotených slov obsiahnutých v tweetoch |
|-----------------|--|
| pred referendum | 21416 |
| po referende | 19584 |

Počet analyzovaných tweetov presiahol počet 10000, pred aj po referende, pozri Tab 2. Zhruba 25 % retweetov sme zaznamenali pred referendum a približne 83 % po ňom, pozri Tab 3. Tweety obsahovali, okrem hľadaného, viacero hashtagov a najfrekvencovanejšie príbuzné hastagy boli #VoteLEAVE, #BrexitNoww a #Euref čo už jasne naznačuje aký výsledok analýzy sme mohli očakávať. Tabuľka 4 ukazuje počet hodnotených slov.

Hodnota sentimentu sa pohybovala okolo -830 bodov, a teda je jasné, že používatelia twitteru sú naklonení na jednu stranu. Bodové ohodnotenie je ale potrebné pochopiť podľa kontextu zadanej otázky. Mínusová hodnota napovedá, že sentiment sa pohybuje v negatívnej rovine, otázkou však je, či to znamená, že používatelia twitteru chcú odchod Británie z EÚ alebo sú práve proti nemu. V tomto rozhodovaní nám môžu pomôcť hashtagy, ale aj náhľad najviac krát retweetnutých príspevkov. Zistili sme teda, že tweety sú ladené prevažne negatívne, kde negatívita smeruje na EÚ a tým pádom volia autori k svojmu príspevku minimálne jeden zo spomínaných hashtagov.



Obr. 1 Vývoj sentimentu pred a po referende

Z Obrázku 1 je patrná výrazná zmena hodnoty sentimentu pred a po referende. Nelíši sa otázka, ktorú sme kládli pri vyhodnocovaní dát, ale mení sa uhol pohľadu. Pred referendom bolo nutné vyhodnotiť negatívny sentiment nie ako reakciu na #brexit, ale ako reakciu na EÚ. Po referende sa môžeme vrátiť k pôvodnej otázke a brať hodnotu sentimentu ako reakciu na #brexit a prebehnuté referendum.

4 Záver

Hlavným cieľom tejto práce bolo vykonanie rozsiahlejšej prípadovej štúdie analyzujúcej tweet z pohľadu celkového sentimentu a jeho vývoja v čase. Vybrali sme veľmi živý hashtag, ktorý súčasne súvisí s bezpečnostnou situáciou hlavne v Európe. Vyhodnotené výsledky a prudká zmena hodnoty sentimentu v kritickom období zlomu, po referende, ukazujú vysokú citlivosť metódy, a teda jej vhodnosť pre získanie hodnôt sentimentu naprieč rozsiahlou skupinou používateľov.

Další výzkum bude zameraný predovšetkým na prepojenie rôznych zdrojov informácií o sentimente z rozdielnych domén pre mnohé záujmové skupiny, a to aj pre jednotlivcov a skupiny používateľov. Očakávame, že výsledok takejto analýzy prinesie zlepšenie analýzy sentimentu súvisiacu s možnými bezpečnostnými rizikami. Túto analýzu budeme aj naďalej spracúvať pre anglicky písané zdroje textov, ale rozšírime ju aj na české a slovenské zdroje.

Podakovanie: Výskum bol podporovaný projektami Technologickej agentúry Českej republiky – TAČR-TF01000091, a grantom SGS 2016/175, VŠB-TU Ostrava.

Literatúra

1. Lars Kai Hansen, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, and Michael Etter. Good friends, bad news - affect and virality in twitter. In *The 2011 International Workshop on Social Computing, Network, and Services (SocialComNet 2011)*. 2011.
2. P. Saloun, M. Hruzik, and I. Zelinka. Sentiment analysis - e-bussines and e-learning common issue. In *Emerging eLearning Technologies and Applications (ICETA), 2013 IEEE 11th International Conference on*, pages 339–343, Oct 2013.
3. Petr Saloun, Adam Ondrejka, and Ivan Zelinka. Similarity of authors' profiles and its usage for reviewers' recommendation. In *9th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2014, Corfu, Greece, November 6-7, 2014*, pages 3–8, 2014.

Annotation:

Twitter and sentiment of #Brexit

#Brexit before and after the UK referendum was incline thousand times a day. History will answer if it was good or bad decision of British people. Serious social network analysis could help us to be in touch with people's sentiment and whit such a knowledge be prepare for the Future. They are social, economical, security circumstances linked and weaved together.

We analyzed more that 10 thousand tweets through 3 months before the referendum, and one thousand tweets after it. Tweets were containing more hash-tags that the mentioned one, such tags were #VoteLEAVE, #BrexitNoww, and #Euref, their meaning was clear, and the sentiment strength was evident and increasing. So, social network analysis evaluated correctly the coming decision. The responsible people, such a politicians and bank-head-quartets have had an opportunity to do their job, does not matter what news and tabloids were talking on their headlines news. Such an sentiment analysis could be useful in the future, as it is serious. The case study and its evaluation is given ere as well.

Aplikácie inteligentných znanostných technológií

Extrakcia štruktúrovaných objektov z webových portálov na pár klikov

Peter Gurský, Milan Vereščák

Ústav informatiky, Prírodovedecká fakulta, Univerzita Pavla Jozefa Šafárika v Košiciach
Jesenná 5, 040 11 Košice, Slovensko

`peter.gursky@upjs.sk`, `mverescak@gmail.com`

Abstrakt. V tomto aplikačnom príspevku predstavíme zásuvný modul do prehliadača, Exago, pomocou ktorého si používateľ pomocou jednoduchých úkonov dokáže anotovať objekty na webovom portáli. V spolupráci s aplikačným serverom, je možné spustiť sťahovanie a následnú extrakciu atribútov týchto objektov do relačnej databázy na ďalšie spracovanie.

Typ príspevku: Aplikačný príspevok

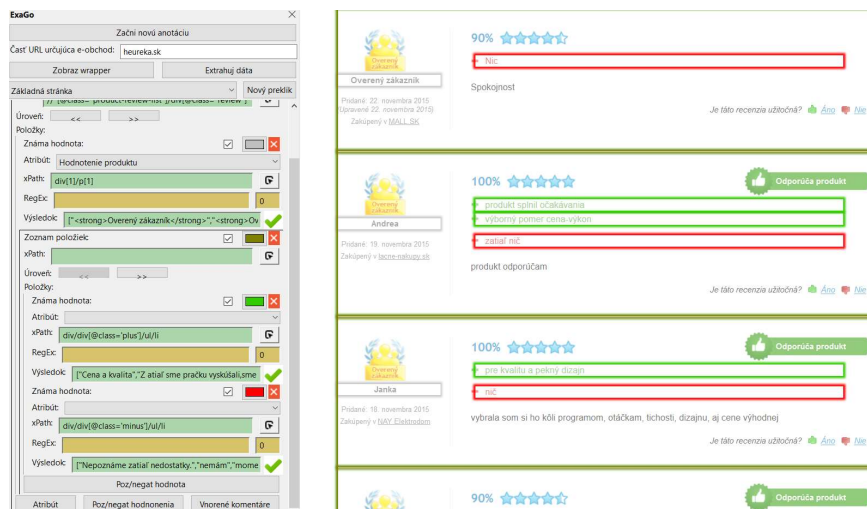
Kľúčové slová: anotácia, extrakcia štruktúrovaných dát z webu, zásuvný modul do prehliadača

1 Úvod

Projekt Kapsa[1] má za cieľ vytvorenie metavyhľadávača produktov, ktorý by umožnil reálne porovnávanie produktov internetových obchodov na základe ich vlastností a komentárov používateľov. Súčasný metavyhľadávač získavajú štruktúrované dáta z internetových obchodov na základe súkromnej komunikácie s týmito obchodmi. V projekte Kapsa sme sa rozhodli pre extrakciu informácií o produktoch priamo z ich webovej prezentácie extrakciou z portálov internetových obchodov, čo umožňuje získanie dát z oveľa väčšieho množstva obchodov a poskytnúť širšiu ponuku produktov, väčšiu vzorku komentárov k produktom, ako aj porovnanie ceny a podmienok väčšieho množstva obchodov.

Na dosiahnutie tohto cieľa je nevyhnutné realizovať pravidelné získavanie všetkých relevantných dát o produktoch z internetových obchodov. Nakoľko každý internetový obchod má vlastný dizajn, vytvorili sme anotačný nástroj, umožňujúci používateľsky nenáročnú anotáciu. Ak chceme extrahovať dáta z nového portálu, alebo v prípade zmeny dizajnu predtým sťahovaného portálu, stačí vytvoriť novú sadu pravidiel na extrakciu za pár minút.

V príspevku prezentujeme anotačný nástroj Exago, ktorý pomáha označovať všetky relevantné časti detailových stránok produktov - atribúty, obrázky, komentáre a špecifikovať pravidlá pre sťahovací a extrakčný server, ktorý následne môže zadaný internetový obchod preliezť a všetky produkty z neho vyextrahovať do databázy.



Obr. 1. Príklad obrazovky pri anotácii komentárov nástrojom Exago.

2 Nástroj Exago

Webových extrakčných systémov je mnoho (ich porovnanie napr. v [2]). Tieto systémy sa delia na *manuálne* [3, 4], ktoré vyžadujú programovanie v nejakom pseudojazyku, *automaticky konštruované extraktory* [5, 6], ktoré vytvoria extrakčný systém na základe kompletnej ručnej anotácie a extrakcie niekoľkých príkladov webových stránok, *automaticky konštruované extraktory s čiastočnou podporou používateľa* [7, 8], ktoré vytvárajú extrakčný systém bez potreby príkladov extrakcie a na *automatické extraktory bez podpory používateľa* [9, 10], ktoré vytvárajú extrakčné systémy analýzou opakujúcich sa vzorov na webových stránkach.

Nástroj Exago je systémom na vytváranie automaticky konštruovaného extraktora s čiastočnou podporou používateľa. Presnejšie, vytára sadu pravidiel, ktorú využíva server na sťahovanie a extrakciu anotovaných dát. Nástroj Exago je naimplementovaný ako doplnok do prehliadača Firefox. To prináša niekoľko výhod. Nástroj je multiplatformový a jednoducho inštalovateľný. Väčšina anotácie sa dá zrealizovať iba udalosťami myši priamo na webovej stránke, ktorú práve anotujeme.

Celá anotácia sa dá vykonať na jednom príklade produktu internetového obchodu a pravidlá, ktoré touto anotáciou vzniknú, sú použiteľné na stiahnutie a extrakciu kompletných dát všetkých produktov daného internetového obchodu. Všetky anotované časti webovej stránky sú okamžite vizuálne farebne označené priamo na webovej stránke, ako je vidieť na obrázku 1.

V rámci anotácie používateľ označuje rôzne typy objektov, ktoré sú pre stránky typu internetových obchodov typické:

- Doménovo nezávislé atribúty ako sú napr. cena, názov, alebo počet kusov na sklade, ktoré majú vo webovej prezentácii všetkých produktov stabilné miesto, prípadne sa vyskytujú ako súčasť URL adresy ako napr. doména alebo identifikátor produktu
- Oblasť doménovo závislých atribútov, ako napr. uhlopriečka displeja, počet otáčok alebo objem, ktoré sa typicky zobrazujú v nejakej tabuľke, alebo zozname ako dvojice názov atribútu a jeho hodnota.
- Zoznam komentárov a ich atribúty, ktoré sú kombináciou predchádzajúcich dvoch typov
- Obrázky
- Prekliky (vrátenie AJAX volaní) napr. na podstránku s komentármi, alebo do galérie obrázkov, ak všetky dáta nie sú iba na jednej detailovej stránke produktu
- Stránkovanie (angl. pagination), pri ktorom je potrebné prejsť niekoľko stránok na nájdenie všetkých hodnôt
- Pravidlá pre sťahovanie portálu, slúžiace na orezanie prehľadávanej časti portálu a identifikáciu detailových stránok

3 Extrakcia dát

Server prijme anotačné pravidlá vo formáte JSON a okrem samotného sťahovania a extrakcie umožňuje aj plánovanie ďalších sťahovaní pre opakované obnovovanie aktuálnosti extrahovaných údajov. Toto plánovanie, ako aj monitorovanie a konfigurácia bežiacich sťahovaní sú realizované cez webové rozhranie.

Samotná extrakcia všetkých produktov z webu je koordináciou dvoch nástrojov – crawler a extraktor. Crawler prechádza všetky relevantné stránky internetového obchodu a v prípade, že pri prechádzaní webových stránok narazí na takú, ktorá spĺňa pravidlá pre detailovú stránku produktu, spustí extraktor daného produktu. Extraktor na základe pravidiel definovaných kombináciou XPath a regulárnych výrazov extrahuje dáta do databázy. Ak sú v pravidlách aj prekliky a stránkovania, tie sa realizujú pomocou nástroja Selenium, ktorý dokáže simulovať akcie používateľa na webe, čo umožňuje aj doťahovanie obsahu cez AJAX volania.

4 Záver

Niektoré metódy odvodzovania pravidiel na základe myšacích udalostí, ktoré Exago využíva, sme už popísali v [11]. Ďalšie metódy, ktoré by ešte viac automatizovali anotáciu sú cieľom nášho ďalšieho skúmania.

Nástroj Exago bol vytvorený na uľahčenie anotácie a extrakcie dát z ľubovoľných internetových obchodov. Je však použiteľný na anotáciu a extrakciu aj iných zoznamov objektov z webových portálov reprezentovaných svojimi atribútmi.

Podakovanie: Túto prácu podporilo MŠVVaŠ SR v rámci projektu VEGA 1/0073/15.

Literatúra

1. Webová stránka projektu Kapsa: <http://kapsa.sk/>
2. B. Liu: Web Data Mining: Exploring Hyperlinks, contents and Using Data, Second edition, Springer 2011. ISBN 978-3-642-19459-7
3. V. Crescenzi, G. Mecca: Grammars have exceptions. Information Systems, 23(8): 539-565, 1998.
4. T. Furche, G. Gottlob, G. Grasso, C. Schallhart, A. Sellers: XPath: A language for scalable data extraction, automation, and crawling on the deep web. The VLDB Journal 22(1): 47-72, 2013
5. C. Hsu, M. Dung: Generating finite-state transducers for semi-structured data extraction from the Web. Information Systems, 1998, 23(8): p. 521-538.
6. Muslea, S. Minton, C. Knoblock: A hierarchical approach to wrapper induction. In Proceedings of Intl. Conf. on Autonomous Agents (AGENTS-1999) 1999.
7. C.-H. Chang, S.-C. Kuo: OLERA: A semi-supervised approach for Web data extraction with visual support. IEEE Intelligent Systems, 19(6):56-64, 2004.
8. Hogue, D. Karger: Thresher: Automating the Unwrapping of Semantic Content from the World Wide Web. Proceedings of the 14th International Conference on World Wide Web (WWW), Japan, pp. 86-95, 2005.
9. V. Crescenzi, G. Mecca, P. Merialdo: RoadRunner: towards automatic data extraction from large Web sites. Proceedings of the 26th International Conference on Very Large Database Systems (VLDB), Rome, Italy, pp. 109-118, 2001.
10. Arasu, H. Garcia-Molina: Extracting structured data from Web pages. Proceedings of the ACM SIGMOD International Conference on Management of Data, San Diego, California, pp. 337- 348, 2003.
11. P.Gurský et al.: Extracting product data from e-shops. Proceedings of ITAT 2014, CEUR Workshop Proceedings Vol. 1214, pp. 40-45, ISSN 1613-0073

Annotation:

Extraction of structured objects from web portals by few clicks.

The paper presents the annotation tool Exago combined with server for crawling and extraction of products from e-shops. With Exago, the add-on for Firefox, a user can annotate product detail page in web browser mostly by clicks. The tool allows annotating attributes, images, comments and can deal with clicks and paginations needed to gain all relevant product data. It is also a configuration tool for crawling to reduce the number of pages to crawl from and identify the product detail pages. The server extracts data of all products from the annotated e-shop to relational database in structured form. The possible clicks are made by Selenium that can handle regular links as well as AJAX calls. Exago tool is created mainly to extract e-shop products' data, but it can be used to extract other lists of objects with similar object-attribute structure as well.

Vizualizace dat s využitím frameworku AngularJS

Petr Kukrál, Martin Dostal, Dalibor Fiala

Katedra informatiky a výpočetní techniky
Západočeská univerzita v Plzni
Univerzitní 8, 306 14 Plzeň, Česká republika

{kukral, madostal, dalfia}@kiv.zcu.cz

Abstrakt. Tento příspěvek se věnuje tvorbě webové aplikace pro vizualizaci dat s využitím JavaScriptu. Klientská část aplikace je implementována ve frameworku Angular od společnosti Google a serverová část je realizována v jazyce PHP. V článku porovnáváme existující možnosti vizualizace dat v oblasti webových technologií a zabýváme se srovnáním nejpoužívanějších JavaScriptových frameworků a knihoven Angular, React a jQuery. Tyto technologie jsou porovnávány jak z hlediska náročnosti implementace, tak z pohledu výkonnosti. Nakonec přinášíme i návod na konverzi pluginu ze starší knihovny jQuery do novějšího Angularu.

Typ příspěvku: Aplikační příspěvek

Klíčová slova: data, vizualizace, JavaScript, Angular, grafy

1 Úvod

S rostoucím množstvím dat roste nutnost tato data správně vizualizovat. Díky vizualizaci si můžeme uvědomit souvislosti mezi daty a vizualizace nám slouží k rychlému přehledu o situaci, který bychom z objemných dat určili jen velmi obtížně. V tomto článku se budeme věnovat vizualizaci dat na webu s použitím moderního frameworku Angular.

Angular [1] je javascriptový framework (odtud také pojmenování AngularJS) vyvíjený společností Google. Stručně si ukážeme jak s jeho pomocí vizualizovat data. Porovnání Angularu s knihovnami React [3] a jQuery [4] je velmi diskutované téma. Proto je zde srovnáme již v konkrétních příkladech vizualizace dat. Ukážeme si klady a zápory jednotlivých technologií a popíšeme si, v jakých projektech je vhodné uvedené technologie použít.

2 AngularJS a jiné technologie

Než se pustíme do samotné implementace vizualizační aplikace v Angularu, musíme vybrat správnou technologii pro zobrazování dat. V práci používáme různé druhy vizualizace dat na webu. Zde si ukážeme (viz tabulku 1), k čemu jsou vhodné, a jaká mají omezení. Konkrétně se podíváme na technologie HTML5 canvas, SVG a HTML elementy.

Tab 6. Porovnání technologií pro vizualizaci dat.

| | can- vas | SVG | HTML a CSS |
|--|-------------|-----|---------------|
| Vykreslení sloupcového grafu | ano | ano | ano |
| Vykreslení koláčového grafu | ano | ano | ne |
| Vykreslení spojnicového grafu | ano | ano | ne |
| JS události se mohou vázat k vykresleným elementům | ne | ano | ano |
| Ovlivňování barev grafu pomocí CSS | ne | ano | ano |
| Pro základní kreslení není nutné používat JavaScript | ne | ano | ano |
| Responzivnost | ne | ano | ano |

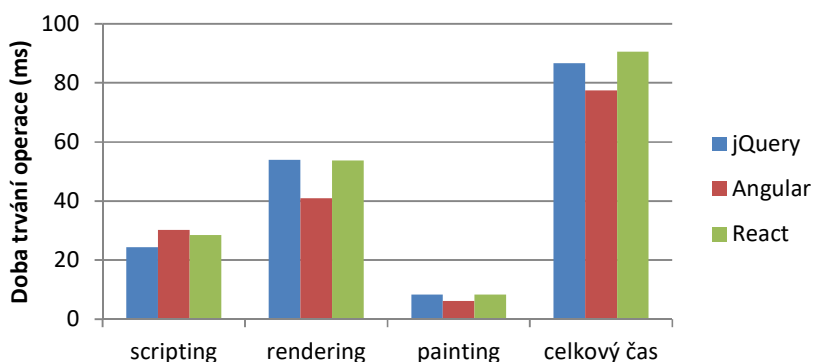
Na Angularu nás na první pohled nejvíce zaujme šablonovací systém. Nicméně Angular toho nabízí mnohem více včetně velkého množství předpřipravených služeb a možností jak naši aplikaci rozšiřovat. Další výhodou Angularu je, že již od začátku nám dává možnost testování. Dalo by se tedy říci, že Angular nás od začátku vede k dobrým návykům vytváření aplikace. JQuery je knihovna, která mimo jiné umožňuje měnit a vykreslovat HTML do stránky, reagovat na události a tvořit animace. React je knihovna, která se zaměřuje na vykreslení a změnu HTML. Přehled některých vlastností Angularu a jeho dvou alternativ je v tabulce 2.

Tab 7. Porovnání některých vlastností Angularu, Reactu a jQuery.

| | Angular | React | jQuery |
|------------------------------------|-----------|----------|----------|
| Práce s HTML DOM | ano | ano | ano |
| Animace | ano | ano | ano |
| Provázání dat s vykreslovaným HTML | ano | ano | ne |
| HTML šablony | ano | ano | ne |
| AJAX | ano | ne | ano |
| RESTful API | ano | ne | ne |
| Podpora MVC architektury | ano | ne | ne |
| Rok vydání | 2009 | 2011 | 2005 |
| Typ | framework | knihovna | knihovna |

3 Dosažené výsledky

V jednom z testů výkonnosti technologií pro vizualizaci dat jsme se zaměřili na měření času přímo v prohlížeči. Výsledné průměrné časy několika fází pěti vykreslování sloupcového grafu jsou zobrazeny na Obr. 1. Test dopadl dle očekávání. JQuery je nejrychlejší při vykonávání skriptu. Angular a React kromě kontroly závislostí navíc porovnávají data a vykreslují pouze ta, která se změnila. JQuery je ale náročné na samotné vykreslení (rendering). To z důvodu, že se překreslují i ty sloupce, u kterých se hodnota nezměnila. Ve výsledných časech je pak nejrychlejší Angular. Naopak nejpomalejší je React. To může být způsobeno tím, že graf v Reactu je do aplikace napojen přes direktivu ngReact, která umožňuje obě technologie propojit. NgReact tak může mít vliv na čas vykonávání skriptu grafu.



Obr. 1. Porovnání jQuery, Angularu a Reactu na úrovni prohlížeče.

V této práci jsme se z části věnovali i přepisování existujících pluginů ze starší technologie jQuery do novější technologie Angular. (Příkladem takového pluginu může být plugin, který vytvoří z existujících dat spojnicový graf.) Máme hned několik možností jak postupovat (viz tabulku 3). Všechny tyto varianty jsou odlišné a ovlivňují výslednou náročnost a kvalitu přepsaného řešení.

Tab 8. Porovnání jednotlivých postupů přepisování pluginu z jQuery do Angularu.

| | Výhody | Nevýhody |
|---|---|---|
| Postupné přepisování | V každé fázi je program spustitelný. Nemusíme předem znát veškeré funkce Angularu | Výsledný kód je zpravidla více závislý na jQuery, než by musel. |
| Přidávání funkcionality do předem vytvořené direktivy | V každé fázi je program spustitelný. Nejsme tak závislí na knihovně jQuery. | Jsou kladeny větší nároky na komplexní znalosti Angularu. |

| | | |
|---|--|---|
| Z jQuery převzít pouze matematické vzorce | Nemusíme být vůbec závislí na knihovně jQuery. Toto řešení má největší potenciál využít nejlepší praktiky vytváření kódu (<i>best practices</i>) z Angularu. | Vysoká náročnost. Komplexní znalost Angularu a postupů na vytváření direktiv. |
|---|--|---|

4 Závěr

Cílem práce zmiňované v tomto článku bylo vytvořit sadu modulů pro vizualizaci dat v Angularu. V práci ukazujeme, jaké technologie je možné použít při vizualizaci na webu, a porovnáváme tři různé knihovny. Také jsme nastínili postup, jak je možné přepsat do Angularu některý z existujících pluginů. V tuto chvíli je vyvíjen Angular 2, avšak aplikaci jsem vytvořili v Angularu 1, neboť Angular 2 je stále ještě v beta verzi. Zdrojové kódy aplikace jsem zveřejnili na portálu Github [2] a uvolnili jsme je i pod nejpožívanějšími licencemi GPL a MIT.

I když byl Angular v testech vykreslování grafu v prohlížeči nejrychlejší, je třeba také přihlídnout k tomu, že se jedná o klientskou aplikaci, která se před spuštěním musí do prohlížeče stáhnout. Je tedy nutné brát v úvahu rozsah (počet znaků) prováděného skriptu. V tomto ohledu je Angular až za Reactem a nejhůře dopadl plugin v jQuery, jehož stažení do prohlížeče kvůli největšímu počtu znaků trvalo nejdelší dobu. Takže dle provedených experimentů je lepší používání jQuery tam, kde vytváříme serverově orientovanou aplikaci, tedy aplikaci, kde větší část včetně renderování HTML vytváříme na serveru a Javascript v tomto případě používáme jen pro drobné úpravy stránky, jako jsou validace a animace. Pokud ale vytváříme jednostránkovou aplikaci, tedy aplikaci, ve které většinu času zůstaneme na jedné stránce, je lepší použít novější technologii Angular, která je ale výkonnostně srovnatelná s Reactem.

Poděkování: Tato publikace byla podpořena projektem LO1506 Ministerstva školství, mládeže a tělovýchovy ČR.

Literatura

1. Kozlowski, P., Darwin P.B.: Mastering Web Application Development with Angular. Packt Publishing, (2013).
2. Kukrál P.: Dashboard Github, [Online; získáno 9. července 2015]. Dostupné z: <https://github.com/Jeriii/op-dashboard>
3. O'Shannessy, P.: React. [Online; získáno 7. února 2016]. Dostupné z: <https://facebook.github.io/react/docs/component-specs.html#lifecycle-methods>
4. Resig J.: JQuery - kuchařka programátora. Computer Press, (2010).

Annotation:

Data Visualization Using AngularJS

This paper deals with the creation of a data visualization web application using JavaScript. The client-side application is implemented in the Angular framework from Google and the server side is developed in the PHP programming language. In the article, the existing possibilities of data visualization in the field of web technologies are compared and we also compare the most widely used JavaScript frameworks and libraries like Angular, React, and jQuery. We assess these technologies from the perspective of implementation requirements as well as the performance point of view. Finally, we also discuss how plugins from the older jQuery library can be converted to the newer Angular framework.

EEG a rozpoznávanie výrazu tváre: Porovnanie prístupov na meranie emócií

Róbert Móro, Jakub Šimko, Peter Gašpar,
Tomáš Matlovič, Jozef Tvarožek, Mária Bieliková

Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava, Slovenská republika

{meno.priezvisko}@stuba.sk

Abstrakt. Informácia o aktuálnej emócii používateľa je cennou spätnou väzbou využiteľnou pri adaptívnom správaní aplikácií ako aj pri post-hoc analýze ich použiteľnosti. Pre spoľahlivé zisťovanie emócií sa však spravidla musíme spoľahnúť na špeciálny hardvér, ktorý nemá dobrú penetráciu a je intruzívny. Kompromisom sa javí byť použitie kamier, sľubné je však aj používanie cenovo dostupných EEG senzorov. V tomto článku prinášame dve štúdie, majúce za cieľ porovnanie existujúcich prístupov merania emócií používateľov.

Typ príspevku: Výskumný príspevok

Kľúčové slová: emócie, rozpoznávanie tváre, EEG, správanie používateľa

1 Úvod: Odhad emócií používateľa je na nezaplatenie

Spoľahlivý odhad aktuálneho emocionálneho stavu používateľa je cennou informáciou. Významnou oblasťou jeho využitia je adaptácia a personalizácia v inteligentných systémoch. V reakcii na informáciu o emocionálnom stave by bolo možné prispôbovať obsah na sociálnej sieti, či nastavovať náročnosť úloh vo vzdelávacom systéme. Ďalšou oblasťou využitia je vyhodnocovanie použiteľnosti softvéru: pri používateľských štúdiách nás výskyt emócie, špeciálne negatívnej, môže rýchlo upozorniť na problematické miesta v rozhraniach a scenároch.

Automatické meranie emócií máme dnes k dispozícii ako produkt v podobe rôznych zariadení a softvéru. Zároveň je však predmetom výskumu, pretože existujúce prístupy a riešenia nie sú ideálne z hľadiska *presnosti*, *neinvazívnosti* a *dostupnosti*. Tieto tri vlastnosti stoja navzájom proti sebe a posilnenie jednej znamená ústupky v druhej. Jeden extrém predstavujú prístupy odhadujúce emócie z tradičných periférnych zariadení (klávesnica, myš) [6] a z odhadov sémantiky používateľských akcií [5, 7]. Keďže nevyžadujú špecializovaný hardvér, sú vysoko dostupné a neinvazívne, no zároveň veľmi nepresné a nie vždy aplikovateľné. Opačným extrémom sú presné, no invazívne nátené senzory fyziológie ľudského tela (merače dýchania, tepu, vodivosti kože) [8], ktoré

majú zároveň nevýhodu malého rozšírenia pre štúdie väčšieho rozsahu. Do tejto kategórie zariadení môžeme zaradiť aj elektroencefalograf (EEG), ktorý sníma elektrické signály z mozgu. Ako kompromis s ohľadom na všetky tri vlastnosti sa javí použitie rozpoznávania tváre pomocou kamier, najmä hlbkových. Kamery nie sú invazívne, často stačia obyčajné webové kamery a aj v prípade špeciálnych hlbkových kamier (ako napr. *Kinect* či *Creative Senz3D*¹) majú určité rozšírenie (najmä vďaka hernému priemyslu). Na druhej strane sú spravidla citlivejšie na vplyv vonkajších činiteľov, akými sú napríklad nevhodné osvetlenie alebo okuliare, ktoré účastníci nosia. S príchodom EEG senzorov *Epoc* od firmy *Emotiv*² (ale aj ďalších, napr. od *Neurosky*³), ktoré sú v porovnaní s klasickým EEG menej invazívne, vyžadujú menšiu obsluhu a sú cenovo dostupné, môžeme hovoriť o pokuse o prienik EEG senzorov do tejto tretej kategórie zariadení. Otázna však zostáva ich presnosť a teda aj možnosť spoľahlivého použitia pre úlohu detekcie a merania emócií.

Metódam merania emocionálneho stavu človeka venujeme pozornosť aj v rámci aktivít v laboratóriách Výskumného centra používateľského zážitku a interakcie (UXI@FIIT, <http://uxi.sk>). Máme k dispozícii zariadenia (spolu s obslužným softvérom), ktoré sú na meranie emocionálneho stavu využiteľné: hlbkové kamery, EEG senzory, okulografy (sledovače pohľadu), fyziologické senzory. Cieľom štúdií prezentovaných v tomto príspevku je experimentálne overenie kvality merania emócií pomocou dostupných zariadení a zhodnotenie možností ich využitia pre úlohy spojené s prispôbovaním a overovaním použiteľnosti.

2 Štúdia 1: Riešenia založené na rozpoznávaní tváre

V prvej štúdií sme sa zamerali na prístupy získavania emócií založené na rozpoznávanie výrazov tváre. Išlo o kvalitatívnu štúdiu s cieľom preskúmať a porovnať možnosti dvoch nástrojov: *Noldus FaceReader*⁴ a *Shore*⁵ (Fraunhofer IIS). Skúmali sme, aké ponúkajú možnosti analýz a automatického vyhodnocovania, a ako sa vedia vysporiadať s negatívnymi činiteľmi.

Účastníkom štúdie sme vo webovom prehliadači premietali sériu 35 obrázkov, ktoré sme vybrali z anotovanej dátovej množiny [2]. Táto dátová množina pôvodne obsahovala 730 obrázkov, ku ktorým bola priradená hodnota náboja (angl. valence) a vybudenia (angl. arousal). Každý z 35 obrázkov sme účastníkovi ukázali na 7 sekúnd, počas ktorých sa mal rozhodnúť o tom, či na neho vplýval pozitívne (1) alebo negatívne (-1), a to na spojitnej stupnici v intervale od -1 po 1. Počas celého experimentu sme zároveň zaznamenávali tvár účastníkov.

Nástroj *Noldus FaceReader* poskytuje analýzu a vizualizáciu emócií z ľudskej tváre. V rámci našich experimentov sme zistili, že nástroj dokáže v relatívne rýchlom čase

¹ <http://us.creative.com/p/web-cameras/creative-senz3d>

² <http://emotiv.com/>

³ <http://neurosky.com/>

⁴ <http://www.noldus.com/human-behavior-research/products/facereader>

⁵ <http://www.iis.fraunhofer.de/en/ff/bsy/tech/bildanalyse/shore-gesichtsdetektion.html>

analyzovať obrázky, zaznamenané videá a živý prenos z kamery počítača. Medzi základné vlastnosti patrilo rozpoznanie šiestich emocionálnych stavov: radosti, smútku, hnevu, prekvapenia, strachu, znechutenia a neutrálneho stavu. Z pokročilých možností môžeme vyzdvihnúť rozpoznanie základných črt účastníka (vek, pohlavie, etnická príslušnosť), podrobných črt tváre (otvorené/zatvorené oči a ústa, prítomnosť brady, fúzov a okuliarov) a tiež polohy tváre. Výhodou nástroja je, že dokáže rozpoznávať emócie na základe štyroch modelov tváre: všeobecný, detské tváre, tváre východoázijských ľudí a tváre starších ľudí. Nástroj sa dokáže spresňovať voliteľnou kalibráciou a vysporadúva sa aj s horšími svetelnými podmienkami, hoci pri protisvetle sme zaznamenali problémy u ľudí nosiacich okuliare.

Obdobným nástrojom určeným najmä pre komerčné účely bol *Shore* (od *Fraunhofer IIS*), ku ktorému sme mali k dispozícii iba demo verziu. Z pohľadu kvality boli však oba softvéry veľmi vyrovnané. Výhodou nástroja *Shore* bola analýza emócií v reálnom čase. Okrem rozpoznávania emócie umožňoval detegovať rôzne veľkosti a otočenia tváre, rozpoznanie očí, úst, pohlavia a veku. Tieto možnosti však boli dostupné samostatne, a nie ucelene ako v prípade nástroja *Noldus FaceReader*. Ďalšou devízou nástroja *FaceReader* bola kvalitnejšia kontinuálna kalibrácia účastníkov. Zároveň nástroj *Shore* neumožňoval vyjadriť percentuálnu mieru detegovanej emócie.

3 Štúdia 2: Riešenie založené na EEG

V druhej štúdii sme sa zamerali na rozpoznávanie emócií s využitím elektroencefalografu (EEG) pomocou nami navrhutej metódy, ktorá využíva metódu podporných vektorov (angl. support vector machines, SVM) na natrénovanie klasifikátora určujúceho jednu zo siedmich emócií (radosť, smútok, znechutenie, hnev, strach, prekvapenie a neutrálnu emóciu, t. j. rovnaké ako *Noldus FaceReader*). Črtami využitými pri klasifikácii sú sila alfa a beta vln vypočítaná z nameraných hodnôt elektrického signálu, hodnoty náboja a vybudenia (angl. valence a arousal) [1] a ich extrémne a priemerné hodnoty pre daný stimul.

Pre overenie našej metódy, ako aj presnosti a spoľahlivosti EEG senzora *Epoc* od firmy *Emotiv* sme zrealizovali štúdiu vychádzajúcu z predchádzajúcej práce v tejto oblasti [4]. Štúdia pozostávala z 20 jednominútových úsekov hudobných videí, ktoré mali v účastníkoch evokovať jednu dominantnú emóciu. Väčšina videí bola prebraná z [4]. Na orchestráciu experimentu sme použili *Tobii Studio*, ktoré umožňovalo zobrazenie videí a k nim prislúchajúcich otázok vo zvolenom poradí; dáta zo sledovania pohľadu získané *Tobii Studio* sme v rámci tejto štúdie nevyhodnocovali. Okrem toho sme nahrávali tváre účastníkov pomocou kamery *Creative Sens3D*; nahrávky sme použili na určenie emócie pomocou rozpoznávania výrazu tváre nástrojom *FaceReader*.

Pred každým videom sa zobrazila na päť sekúnd čierna obrazovka s bielym fixačným krížom uprostred, na ktorý sa mal účastník pozerieť. Po každom takomto jednominútovom videu nasledovalo vyplnenie dotazníka, ktorý pozostával z troch otázok: (i) „Aká silná bola emócia, ktorú ste pociťovali?“, (ii) „Aká pozitívna bola emócia, ktorú ste pociťovali?“, (iii) „Aká emócia u vás prevládala najviac?“. Na prvé dve otázky účas-

tníci odpovedali zvolením hodnoty od 1 do 10, pričom zvolené hodnoty indikovali subjektívne hodnotenie náboja a vybudenia účastníka. Pri poslednej otázke si vyberali jednu zo siedmich vyššie uvedených emócií.

Štúdiu sme zrealizovali vo výskumnom UXI centre na FIIT s 9 účastníkmi, pričom jedno sedenie trvalo približne 40 minút. Celkovo sme tak nazbierali dátovú sadu 180 emóciou ohodnotených videí, ktorú sme využili na experimentálne overenie nami navrhnutej metódy. Pri využití 5-násobnej krížovej validácie sa nám podarilo dosiahnuť priemernú úspešnosť určenia správnej emócie na testovacej množine 58% so štandardnou odchýlkou $\pm 6\%$. Museli sme sa pritom vysporiadať s nevyváženosťou dátových vzoriek, keď niektoré emócie (napr. hnev alebo strach) boli v dátovej sade zastúpené len minimálne; za týmto účelom sme využili nadzorkovanie počas fázy tréningovania. Výsledky sme tiež porovnali s určením emócie rozpoznávaním výrazov tváre pomocou nástroja *Noldus FaceReader*. Tento nástroj pre daný časový moment neurčuje jednu konkrétnu emóciu, ale pomer emócií. Pri započítaní len dominantnej emócie sme dosiahli úspešnosť 19%, čo je výrazne menej ako pomocou EEG senzora.

4 Pripravenosť technológií pre ich využitie vo výskume a praxi

Prvá realizovaná štúdia bola síce len kvalitatívna s cieľom preskúmať možnosti existujúcich nástrojov, výsledky rozpoznávania emócií pre prezentované stimuly (obrázky) však boli povzbudivé, aj keď by si vyžadovali pre potvrdenie rozsiahlejší experiment. Na druhej strane, na úlohe rozpoznávania emócií pri pozeraní hudobných videí dosiahol prístup využívajúci detekciu výrazov tváre podstatne horšie výsledky ako EEG. Môže to byť spôsobené jednak samotným použitím EEG senzora, ktoré mohlo v účastníkoch vyvolať istú strnulosť, ale aj charakterom prezentovaných videí, pri ktorých bola evokovaná emócia zrejme slabšia. V budúcnosti by tak zrejme bolo dobré nebrať do úvahy len dominantnú emóciu, ale aj ďalšie detegované z výrazu tváre, keďže aj pri prirodzených stimuloch (nezameraných na vyvolanie konkrétnej emócie; napr. aplikácia, s ktorou pracuje používateľ) je predpoklad, že prejavovaná emócia bude slabšia.

Pri EEG senzoroach sa ukázalo, že hoci sú už cenovo prístupné, stále vyžadujú relatívne zdĺhavú fázu prípravy (napr. vlhčenie elektród), čo zatiaľ zamedzuje ich väčšiemu rozšíreniu v tejto oblasti. Zrealizovali sme malý experiment (na troch účastníkoch) aj s jednoduchším zariadením *Insight* od firmy *Emotiv*, ktorý by mal túto bariéru prekonať (za cenu menšieho počtu elektród), ale jeho spoľahlivosť a presnosť sa ukázali pre daný typ úlohy nepostačujúce. Budúcnosť tak zrejme spočíva v ďalšom rozvíjaní senzorov, aby boli čo najmenej invazívne a zároveň dostatočne spoľahlivé a presné, kde sa zaujímavým javí byť okulometer (sledovač pohľadu). Tento dokáže merať veľkosť zreničky, ktorá indikuje emočné vybudenie človeka [3]. Ďalší perspektívny smer predstavuje kombinácia rôznych prístupov.

Podakovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektov APVV-15-0508, VG 1/0646/15 a KEGA 009STU-4/2014.

Literatúra

1. Blaiech, H., Neji, M., Wali, A., Alimi, A.M.: Emotion Recognition by Analysis of EEG Signals. In: HIS '13: Proc. of 13th Int. Conf. on Hybrid Intelligent Systems, IEEE, (2013), pp. 312–318.
2. Dan-Glauser, E., Scherer, K.: The Geneva Affective Picture Database (GAPED): A New 730-picture Database Focusing on Valence and Normative significance. *Behavior Research Methods*, (2011), vol. 43, no. 2, pp. 468–477.
3. Juhaniak, T., Hlaváč, P., Móro, R., Šimko, J., Bielíková, M.: Pupillary Response: Removing Screen Luminosity Effects for Clearer Implicit Feedback. In: UMAP 2016: Posters, Demos, Late-breaking Results and Workshop Proc. of the 24rd Conf. on User Modeling, Adaptation, and Personalization. CEUR-WS, (2016), p. 2.
4. Koelstra, S. et al.: DEAP: A Database for Emotion Analysis; Using Physiological Signals. *IEEE Transactions on Affective Computing*, (2012), vol. 3, no. 1, pp.18–31.
5. Korenek, P., Šimko, M.: Sentiment Analysis on Microblog Utilizing Appraisal Theory. *World Wide Web*, (2014), vol. 17, no. 4, pp. 847–867.
6. Kolakowska, A.: A Review of Emotion Recognition Methods Based on Keystroke Dynamics and Mouse Movements. In HSI '13: Proc. of 6th Int. Conf. on Human System Interactions, IEEE, (2013), pp. 548–555.
7. Ortigosa, A., Martín, J.M., Carro, R.M.: Sentiment Analysis in Facebook and its Application to E-learning. *Computers in Human Behavior*, (2014), vol. 31, no. 1, pp. 527–541.
8. Takahashi, K.: Remarks on Emotion Recognition from Bio-Potential Signals. In: Proc. of the 2nd Int. Conf. on Autonomous Robots and Agents, IEEE, (2004), pp. 186–191.

Annotation:

EEG and Face Recognition: Comparison of Approaches for Emotion Detection

Information on the current user emotion is a valuable feedback that can be used for adaptation of the behaviour of applications as well as for post-hoc analysis of their usability. In order to obtain a reliable emotion detection, we usually have to rely on the specialised hardware, which has low penetration and is intrusive. Using web cameras seems a compromise, promising is also the use of affordable EEG sensors. In the paper we present two studies that aim to compare the existing approaches of user emotion detection.

Considerations about Data Processing, Machine Learning, HPC, Apache Spark and GPU

Giang Nguyen, Ján Astaloš, Ladislav Hluchý

Department of Parallel and Distributed Information Processing
Institute of Informatics, Slovak Academy of Science
Dúbravská cesta 9, 845 07 Bratislava, Slovakia

{giang, astalos, hluchy}.ui@savba.sk

Abstract. Recently, the terms Internet of Things (IoT), Big Data and Machine Learning become very hot topics in both research and commercial spheres. IoT refers to the world of devices connected to the Internet, which is the way the massive amount of data is continuously collected, concentrated and managed. Raw data can also come from other processes such as information retrieval, web monitoring, database systems and so on. Mining in such data means of analysis in order to obtain usable results and/or knowledge. This paper presents several considerations about large-scale data, data processing and data mining using machine learning techniques with technological backgrounds towards high performance computing (HPC), Apache Spark and GPU that enable and accelerate the whole process.

Contribution type: Work-in-progress paper

Keywords: data processing, data mining, machine learning, HPC, Spark, GPU

1 Introduction

It is clear that machine learning (ML) algorithms learn from data and data is de facto the heart of many solutions. The availability of high performance infrastructures, technologies and available machine learning libraries in combinations with computational and/or data intensive strategies open nearly unlimited possibilities for data mining (DM). However, one important point is the flexibility of a solution design, which must be done around, at least, the 3Vs (*Volume, Velocity and Variety*) of data towards efficiency criterions such as resources, performance, cost efficiency, etc. A universal solution for the “*Big Data*” challenges does still not exist, however the coupling of strategies and technologies upon mathematical backgrounds and data-centric approach based on real requirements is a good starting point. In practical scenarios with big and large-scale data contexts, the use of incremental algorithms is visibly increased [4][6] with satisfied reported results of models’ performance in comparisons with traditional in-memory algorithms.

2 Data mining using machine learning techniques

Nowadays, the global data production is continually increased by worldwide distributed ubiquitous sensors for long-term monitoring. Mining in such data means of analysis in order to obtain usable results and/or knowledge. Currently, ML techniques in general and supervised learning approaches in particular, play the central role in many practical/commercial cases. In general, ML approaches can be divided [1] into:

- Traditional in-memory learning (offline learning) where whole data for training can be loaded into machine memory. The main advantage of this approach is in many existing algorithms, number of available libraries, each with numerous methods and implementation improvements to achieve precise results. The disadvantage is the memory limitations that imply only use of small data sets.
- Incremental learning (online learning) does not require the whole data to be loaded into the machine memory at once. Instead, it loads the data in batches. These algorithms use limited memory and limited processing time per item, therefore, the input data set can be large-scale without memory limitation. On the other hand, the number of available algorithms are limited in comparison to in-memory approach.
- Distributed learning: which is typically coupled with infrastructure i.e. DAS (Data Analytics Supercomputer e.g. Apache Spark [2]). It is usually applied on very large data sets, which do not fit into memory of one machine. DAS is usually utilized also as a whole ecosystem with data processing, data integration and data management.

If a set of ready for use machine learning methods is extensive, their implementations are also rich and available in many languages with many versions and improvements. The most well-known ML libraries (or collections) are (Tab 1.):

Tab 1. The most well-known ML libraries

| Library (impl. language) | Strong points | Weak points |
|---------------------------|---|--------------------------------|
| Weka3 (Java) | general purpose, GUI, popular | small datasets, GUI, popular |
| MOA (Weka related) | data stream mining, concept drift, recommender systems | |
| R, Python (and libraries) | statistics, ML, very popular | R vs. Python |
| RapidMiner | general purpose, DB connection, popular | |
| Scikit-Learn (Python) | general purpose, popular | small datasets |
| NLTK (Python) Clojure | general purpose, natural language toolkit and text mining | small datasets |
| PyBrain (Python) | neural network, reinforcement learning, evolution, easy use | good for study and experiments |
| MLLib (Scala, Java) | Spark distributed scalable ML framework, growing community | coupled with infrastructure |
| Mahout (Java) | Hadoop ML framework | come with Hadoop overhead |

| | | |
|---------------------------------|--|------------------------------|
| H2O.ai | massively scalable Big Data analysis, distributed processing (Hadoop, Spark) | |
| Shogun (C++) | general purpose, designed for large scale learning, kernel methods, SVM, HMM | |
| LIBSVM (C++) LIBLINEAR (C++) | integrated software, large-scale data | narrowed approach |
| Vowpal Wabbit (C++) | fast out-of-core ML system, on-line learning | limited number of algorithms |
| XGBoost | parallelized general purpose gradient boosting library | narrowed approach |
| MatLab, GNU Octave | scientific libraries | math oriented |

One of the most used *data mining concept and methodology* [8] is CRISP-DM (Cross-Industry Process for Data Mining), which consists of six steps: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. The Data Preparation step consists of sub-steps: Data Transformation, Exploratory Data Analysis (EDA) and Feature Engineering. The group of the first five steps are also called the development phase. The deployment step is also called the production phase.

Although the main interest upon DM/ML is broadly paid to the Modeling step and algorithms, one important point remains the fact that ML algorithms learn from data. Therefore, in practice, Data Understanding and Data Preparation can consume up to 80% of the entire time of every DM using ML techniques project. Data Preparation is also slangy labeled Data Munging or Data Wrangling, which refer to strenuous work. Certain problem-solving techniques e.g. Forward Selection, Backward Eliminations in the Feature Engineering sub-step or grid-search in the Modeling step can lead to computational intensive tasks especially when ML input data is large-scale or big. HPC (high-performance computing) cluster can be utilized for concurrent training of models in order to shorten the development time.

In the following parts, some practical notes around data processing and DM process using ML techniques for commercial and research applications with IISAS participation in recent years are presented.

Malicious behavior detection in mobile devices log domain. When everyone owns and uses mobile devices such as smartphones and/or tablets, the demand of *cybersecurity* and *situational awareness* is pushing towards. This involved work was a part of the six-month pilot research done for IBM Slovakia. The interest was if it is possible to detect malicious behaviors of mobile devices based on collected logs of mobile devices. Raw data - logs from mobile devices belongs to human-generated data class, which are not so “Big” as machine-generated data. Data mining using ML techniques in this domain involved through following obstacles:

- Collected raw logs are extremely noisy for the specific detection purpose. The logs contain a lot of information about continuous monitoring processes such as timing (clocks, alarms, calendars), positions, accelerators, display setting and adapting, network and power monitoring, scanning processes, etc.

- Low occurrences of malicious behaviors - malware related activities, which caused imbalanced classes of data used for supervised ML;
- Feature extraction for data with evolving characteristics i.e. number of applications on mobile devices is changed based on users' demands without any limitations;
- Privacy preserving data mining of personal sensitive information.
- DM process required thorough Data Understanding in collaboration with domain experts, Data Preparation (especially EDA) and Feature Engineering. ML technique applied in this case was simple supervised binary classification with incremental learning. The obtained results were highly satisfied to distinguish malicious behavior from the normal one.

Click-through-rate advertising: raw and ML data are really big in both development and production phases. Applied analyzing techniques are e.g. reservoir (sub)sampling, biases monitoring, smoothing, sliding windows with settable size, forgetting mechanism, etc. came with *adaptive online learning* (retraining in combination with incremental adaptation). ML data is highly imbalanced as usually in many commercial cases that implies boosting one class against the second by reducing number of negative examples. Feature selections and feature combinations are also utilized to improve models' performance. The production infrastructure is high-performance Hadoop cluster of the Magnetic Media Online, Inc. technology company (USA).

Power utility for functional awareness of monitoring stations: raw input data in this case is quite interesting, it is not "Big" in any one of 3Vs, but contains pure numerical and structured data collected from monitoring stations during several years. Such data can be called large-scale, which causes computational intensive tasks with memory consumption in the development phase. The question was if it is possible to realize the production on single machine with limited memory due to cost and energy efficiency. The solution can be any of traditional in-memory approach in a machine with larger memory, incremental learning or distributed learning with Spark installation in single machine for the production. However, the use of the incremental learning to overcome machine memory limitation can be the less painful way on both phases.

3 Machine learning and many-core accelerators

In the recent years the accelerators have been successfully used (not only) in machine learning and deep learning applications [4]. Manufacturers often offer the possibility to enhance hardware configuration with many-core accelerators to improve machine/cluster performance. If we look at the list of top 500 most powerful supercomputers, we can see the increasing trend in both number of systems that employ the accelerators and their performance share. Most popular models of accelerators are based on *MIC (Many Integrated Cores)* and *GPU (Graphics Processing Unit)* architectures. The accelerators are able to offer significant performance increase for many application domains e.g. the work [5] realized in collaboration between TUKE (Technical University of Košice) and IISAS (Institute of Informatics, Slovak Academy of Science). The main feature of the many-core accelerators is massively parallel architecture (e.g. new NVIDIA P100 accelerator contains 3840 CUDA cores), allowing them to speed up computations that

involve matrix-based operations, which is a heart of many ML implementations. Many popular ML frameworks and libraries already offer the possibility to use GPU accelerators to speed up learning process with supported interfaces in various languages e.g.:

Tab 2. Popular ML frameworks and libraries

| Library (impl. language) | Main purposes |
|------------------------------------|--|
| Theano (Python) | math expression compiler |
| Tensorflow (C++, Python) | numerical computation library by data flow graphs |
| Keras (Python) | minimalist, highly modular neural networks library capable of running on top of TensorFlow or Theano |
| Caffe (C/C++, Python, MatLab, CLI) | deep learning framework for image processing |
| CNTK (C++, CLI) | unified deep-learning toolkit that implements CNN and RNN training for speech, image and text data |
| DL4J (Java, Scala) | distributed deep-learning library written for Java and Scala, integrated with Hadoop and Spark |
| Neon (Python) | Nervana's Python-based deep learning library |
| Torch (C/LuaJIT) | NN and optimization libraries that puts GPUs first |
| MatConvNet | Convolutional Neural Networks (CNNs) for MatLab |

Some of them also allow to use optimized CUDA Deep Neural Network (cuDNN) library to improve the performance even further. Similar to the ML libraries mentioned in Section 2, ML libraries with GPU support are also diverted in various implementation levels for various specific purposes such as image, voice and text processing.

The demand for even more powerful hardware for deep learning applications caused that main manufacturer of GPU accelerators NVIDIA made considerable investments to the development of the new architecture called Pascal and special purpose system DGX-1 optimized for many-layered DNN. Among the new features most notable are the „*half-precision*“, which allows to reach 21.2 Teraflops and 160 GB/s bidirectional interconnect that significantly improves the scalability in multi-GPU systems.

The matrix-based operations on Apache Spark can be computationally accelerated under same logic like GPU/CUDA acceleration. Here is a *similar logic* between Apache Spark vs. GPU processing (not only) from ML viewpoint:

- If data fits into memory of one machine, GPU is faster, otherwise Spark;
- Spark logic is similar to CUDA host logic in the mean of SIMD processing;
- Spark network overhead vs. PCI-express transfer overhead;
- MapPartitions is like kernel launch, partitions are like CUDA blocks;
- Model parallelism vs. data parallelism: Data parallelism presents single instruction to multiple data items, ideal workload for a SIMD computer architecture; Model parallelism gives every processor the same data but applies a different model to it; Hybrid approach presents combination of data and model parallelism.

Potential benefits^{1,2} of using GPUs to further accelerate Spark performance is also done with positive results.

4 Conclusions

This paper presents a few considerations about working and mining in large-scale data using ML techniques in our department in recent years. We hope that such notes are useful for readers with nearby research interests and would like to thank to colleagues and reviewers for consultations and advices on the paper preparation.

Acknowledgements: This work is supported by projects VEGA 2/0167/16 and EGI-Engage EU H2020-654142. Simulations and technical realization are realized on the hardware equipment obtained within the project SIVVP ERDF ITMS 26230120002.

Remarks: In addition to standard HPC computational power, SIVVP³ (Slovak Infrastructure for High Performance Computing) HPC clusters are also enhanced by GPU accelerators NVIDIA M2050/M2070 (448 CUDA cores) and K20 (2496 CUDA cores) to allow researchers from Slovakia to use GPU accelerated systems for research purposes. The installed GPU capacity is as follows: Institute of Informatics SAS, Bratislava: 16x K20 + 2x M2070; Matej Bel University, Banská Bystrica: 6x K20 + 2x M2070; Technical University of Košice: 2x K20 + 2x M2070; Institute of Experimental Physics, Košice: 10x K20 + 32x M2070; University of Žilina: 2x M2070; Slovak University of Technology in Bratislava: 8x M2050.

References

1. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A.: A survey on concept drift adaptation. ACM Computing Surveys (CSUR), 2014 Apr 1;46(4):44 pages.
2. Karau H., Konwinski A., Wendell P., Zaharia M.: Learning Spark. Published by O'Reilly Media, Inc. © 2015 Databricks, 242 pages, ISBN: 978-1-449-35862-4.
3. Lacey G., Taylor G. W., & Areibi, S. (2016). Deep Learning on FPGAs: Past, Present, and Future. arXiv preprint arXiv:1602.04283.
4. Lopes N., Ribeiro B.: Machine Learning for Adaptive Many-Core Machines - A Practical Approach. Studies in Big Data, Volume 7, Springer International Publishing Switzerland, 2015, 251 pages, ISBN 978-3-319-06937-1, ISSN 2197-6503.
5. Naščák D., Košťál I., Mikula J., Oljaj A., Astaloš J.: Acceleration of simulation models for raw materials thermal treatment. 12th International Carpathian Control Conference: ICCCC'2011, pp. 207-212, ISBN 978-161284359-9.
6. Rozinajová V. et al: Otvorené smery výskumu v oblasti dátovej analytiky. WIKT 2015 proceedings, pp.4-7, ISBN 978-80-553-2271-1.

¹ <http://www.slideshare.net/continuumio/gpu-computing-with-apache-spark-and-python>

² <http://www.nextplatform.com/2016/02/24/hadoop-spark-deep-learning-mesh-on-single-gpu-cluster/>

³ SIVVP - Slovak Infrastructure for High Performance Computing (<http://www.sivvp.sk/>)

7. Sumeet Dua and Xian Du: Data Mining and Machine Learning in Cybersecurity. CRC Press, Taylor & Francis Group, 248 pages, 2011, ISBN-13 978-1-4398-3943-0.
8. Vadovský, M., Michalik, P., Zolotová, I. and Paralič, J.: Better IT services by means of data mining. IEEE Int. Symposium on Applied Machine Intelligence and Informatics SAMI 2016, pp. 187-192, 2016, ISBN 978-146738740-8.

Podpora zdieľania znalostí vo vzdelávacích kurzoch prostredníctvom CQA systému Askalot

Ivan Srba, Mária Bieliková

Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava, Slovenská republika

{ivan.srba, maria.bielikova}@stuba.sk

Abstrakt. Úspešnosť systémov pre odpovedanie otázok v komunitách (angl. Community Question Answering - CQA) na otvorenom webe (napr. Stack Overflow) viedla k ich aplikovaniu v nových kontextoch (napr. vo vzdelávaní) a v nových prostrediach (napr. v rámci organizácií). Predstaviteľom tohto trendu je aj vzdelávací univerzitný CQA systém Askalot vyvíjaný na FIIT STU. Aby sme plne preskúmali potenciál CQA systémov v edukačnej doméne, nadviazali sme spoluprácu s výskumníkmi z Harvardovej univerzity s cieľom upraviť Askalot ako rozšírenie do MOOC systému edX. Zároveň pracujeme na nasadení Askalotu na ďalších univerzitách v Lugane a v Novom Sade. V príspevku opisujeme návrhové a implementačné riešenia, ktoré poskytli potrebnú flexibilitu a škálovateľnosť pre tieto rôzne prostredia. Zároveň predstavíme, aké výskumné možnosti poskytuje nasadenie Askalotu v rámci domény vzdelávania.

Typ príspevku: Aplikačný príspevok

Kľúčové slová: CQA, MOOC, Askalot, zdieľanie znalostí, komunity študentov

1 Úvod

Od vzniku systémov pre odpovedanie na otázky v komunitách (angl. Community Question Answering – CQA) sa stali tieto systémy významným zdrojom znalostí v priestore súčasného webu. V najpopulárnejších CQA systémoch, ako je napr. Stack Overflow alebo Yahoo! Answers, komunity pozostávajúce z miliónov používateľov zdieľajú svoje znalosti prostredníctvom pýtania sa otázok a poskytovania odpovedí. V poslednej dobe úspešnosť a popularita CQA systémov na otvorenom webe motivuje výskumnú ako aj komerčnú sféru pre ich používanie aj v ďalších oblastiach. V prvom rade bol potenciál CQA systémov rozpoznávaný nielen v kontexte webu, ale aj v doméne vzdelávania [1] alebo v centrách zákazníckej podpory [3]. V druhom rade koncepty CQA systémov nemusia byť využívané len veľkými otvorenými komunitami, ale aj uzavretými skupinami používateľov v rámci organizácií (napr. v sociálnej platforme

IBM Connect [2]). Adaptovanie CQA konceptov v týchto nových oblastiach však prináša nové problémy a výskumné výzvy (napr. ako prispôsobiť funkcionality CQA systémov špecifikám konkrétneho prostredia).

V našej predchádzajúcej práci sme identifikovali nový koncept vzdelávacieho a organizačného CQA systému, ktorý sme následne zrealizovali a overili formou systému Askalot [4]. V kontraste so štandardnými otvorenými CQA systémami (napr. Stack Overflow), systém Askalot¹ zohľadňuje špecifiká vzdelávania (napr. prítomnosť učiteľa, výrazne odlišná úroveň znalostí používateľov) a organizačného prostredia (napr. menšia veľkosť komunity, známosť používateľov). Systém Askalot je zrealizovaný ako webová aplikácia s otvoreným zdrojovým kódom². Askalot je aktuálne nasadený na Fakulte informatiky a informačných technológií Slovenskej technickej univerzity v Bratislave. Zahŕňa komunitu viac ako 1100 študentov a učiteľov, ktorí doteraz poskytli viac ako 560 odpovedí na viac ako 430 otázok.

Na základe dosiahnutých pozitívnych výsledkov sme v uplynulom období nadviazali spoluprácu s:

1. **Harvardovou univerzitou** s cieľom využívať Askalot ako náhradu štandardnej diskusie v MOOC (angl. Massive Open Online Courses) systéme edX.
2. **Univerzitami v Lugane a v Novom Sade** v rámci kooperačného projektu programu SCOPES s cieľom nasadiť Askalot na týchto univerzitách.

Pôvodný návrh systému Askalot (opísaný v príspevku [4]) však bol navrhnutý špecificky pre našu univerzitu a neposkytoval tak potrebnú flexibilitu a škálovateľnosť pre tieto rozličné vzdelávacie prostredia. Dôsledkom toho sme museli jeho dizajn prepracovať a výsledkom je niekoľko dizajnových odporúčaní, ktoré predstavíme v nasledujúcej časti príspevku.

2 Návrh systému Askalot pre rôznorodé vzdelávacie prostredie

Pokým niektoré koncepty a funkcie CQA systémov využívaných v špecifických prostrediach sú prirodzene flexibilné a škálovateľné, niektoré vyžadujú pri širšom nasadení viaceré návrhové a implementačné úpravy. Zmeny vykonané v systéme Askalot sme rozdelili do štyroch skupín.

Modulárna architektúra. Nasledujúce požiadavky a špecifiká rôznych prostredí sme identifikovali dve hlavné konfigurácie systému Askalot, ktoré sme kódovo označili ako *Askalot @university* a *Askalot @mooc*. Následne sme vytvorili tri moduly. Do prvého sme vyčlenili spoločnú funkcionality pre všetky prostredia (napr. vkladanie otázok a odpovedí alebo zoznamy používateľov, kategórií, atď.). Ostatné dva moduly dedia všetky funkcie z tohto primárneho modulu a pridávajú špecifické funkcie pre univerzitné, resp. MOOC prostredie.

Flexibilná integrácia manažmentu používateľov. Askalot poskytuje niekoľko spôsobov či už automatickej alebo manuálnej registrácie a autentifikácie používateľov.

¹ Demo systému Askalot je dostupné na <https://askalot.fiit.stuba.sk/demo>

² Zdrojový kód systému Askalot je dostupný na <https://github.com/AskalotCQA/askalot>

Predovšetkým je možné využiť LDAP autentifikáciu, ktorá je dostupná na mnohých univerzitách. Zároveň Askalot podporuje LTI protokol (angl. Learning Tool Interoperability), ktorý bol špecificky navrhnutý pre výmenu informácií medzi vzdelávacími systémami (vrátane informácií o samotných študentoch). Poslednou možnosťou sú nezávislé používateľské účty priamo v systéme Askalot. S využitím tohto spôsobu autentifikácie je možné nakonfigurovať Askalot tak, že jednotlivé účty môžu byť plne anonymné, čo je dôležité najmä v prípadoch, keď študenti sa odmietajú pýtať otázky pokým je ich identita verejná.

Adaptívna a samo udržiavajúca sa organizácia obsahu. Askalot poskytuje dvojúrovňovú štruktúru obsahu. Na prvej úrovni môže pýtajúci sa používateľ zaradiť svoju otázku do hierarchie kategórií (tie reflektujú formálnu štruktúru vzdelávania, napr. predmety alebo sekcie online kurzu). Na druhej úrovni je možné upresniť tému otázky s využitím značiek (angl. tags). Hierarchia kategórií je v systéme Askalot navrhnutá tak, že zohľadňuje ich pravidelné opakovanie (cez akademické roky alebo opätovného otvárania MOOC kurzov). Následne je možné zvoliť napr. v ktorých kategóriách sa má zobrazovať aj obsah z predchádzajúcich iterácií toho istého predmetu/kurzu.

Široko dostupný prehľad aktivít a notifikácií. V neposlednom rade Askalot poskytuje viacero možností, ako informovať študentov a ich učiteľov o aktivite. Sú to predovšetkým notifikácie zobrazované priamo v systéme, ale aj sumárny email s aktivitou za posledných 24 hodín. Navyše je možné prepojiť Askalot so sociálnou sieťou Facebook a notifikácie sú následne zobrazované priamo v tejto sociálnej službe.

3 Záver a ďalšia práca

Na základe úpravy návrhu a implementácie systému Askalot sme ukázali niekoľko návrhových odporúčaní, ako môžu byť CQA systémy aplikované nielen na webe, ale aj v konkrétnej doméne a prostredí, pričom bolo možné dosiahnuť vysokú úroveň flexibility a škálovateľnosti. To nám v konečnom dôsledku umožňuje nasadiť systém Askalot na viacerých univerzitách ako aj v MOOC systéme edX, kde sa Askalot používa od začiatku septembra 2016 ako súčasť kurzu *QuCryptox Quantum cryptography*³ s celkovým počtom viac ako 5200 zapísaných študentov.

Takéto nasadenie pritom poskytuje významný výskumný potenciál. Predovšetkým sme sa zatiaľ v systéme Askalot sústredili prevažne na prispôbenie základných funkcií a konceptov CQA systémov. Sme si ale vedomí, že aj metódy pre podporu spolupráce počas procesu odpovedania na otázky podliehajú vplyvom špecifik vzdelávacieho a organizačného prostredia. Pokým sa týmito metódami v štandardných otvorených CQA systémoch venuje dostatok pozornosti, v doménovo špecifických CQA systémoch môžeme aktuálne vidieť len prvé výskumné príspevky, napr. pri odporúčaní nových otázok používateľom, ktorí sú vhodnými kandidátmi na poskytnutie odpovede (angl. Question routing) [7].

Navyše doménovo špecifické CQA poskytujú príležitosť overovať výskumné metódy v živých experimentoch. Doteraz v oblasti výskumu CQA systémov boli takéto

³ <https://courses.edx.org/courses/course-v1:CaltechDelftX+QuCryptox+3T2016/info>

živé experimenty len veľmi zriedkavé. Na základe nášho predchádzajúceho komplexného prehľadu prístupov pre podporu spolupráce v CQA systémoch [6] sme zistili, že len 3 zo 169 prístupov boli overené v živých experimentoch. Askalot navyše poskytuje experimentálnu infraštruktúru [5], ktorá umožňuje jednoduché prepojenie syntetických experimentov na dátových sadách (zo systému Askalot ako aj z CQA systémov založených na platforme Stack Exchange) a online experimentov v systéme samotnom.

Pod'akovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektov KEGA 009STU-4/2014 a je čiastočným výsledkom spolupráce v rámci projektu SCOPES JRP/IP, No. 160480/2015.

Literatúra

1. Aritajati, C., Narayanan, N.H.: Facilitating Students' Collaboration and Learning in a Question and Answer System. In: Proc. of the 2013 conf. on Computer supported cooperative work companion - CSCW '13. ACM Press, (2013), pp. 101–106.
2. Luo, L., Wang, F., Zhou, M.X., Pan, Y., Chen, H.: Who Have Got Answers? Growing the Pool of Answerers in a Smart Enterprise Social QA System. In: Proceedings of the 19th int. Conf. on Intelligent User Interfaces - IUI '14. ACM Press, (2014), pp. 7–16.
3. Piccardi, T., Convertino, G., Zancanaro, M., Wang, J., Archambeau, C.: Towards Crowd-based Customer Service: A Mixed-Initiative Tool for Managing Q&A Sites. In: Proc. of ACM conf. on Hum. factors in comp. syst. - CHI '14. ACM Press, (2014), pp. 2725–2734.
4. Srba, I., Bielikova, M.: Askalot: Community Question Answering as a Means for Knowledge Sharing in an Educational Organization. In: Proc. of the 18th ACM Conf. Companion on Computer Supported Cooperative Work & Social Computing - CSCW'15. ACM Press, (2015), pp. 179–182.
5. Srba, I., Bielikova, M.: Design of CQA Systems for Flexible and Scalable Deployment and Evaluation. In: Proceedings of the 16th Int. Conf. on Web Engineering - ICWE '16. Springer, (2016), pp. 439–447.
6. Srba, I., Bielikova, M.: A Comprehensive Survey and Classification of Approaches for Community Question Answering. ACM Trans. Web. 10 (3), 1–63 (2016).
7. Yang, D., Adamson, D., Rosé, C.P.: Question Recommendation with Constraints for Massive Open Online Courses. In: Proc. of the 8th ACM Conf. on Recommender systems - RecSys '14. ACM Press, (2014), pp. 49–56.

Annotation:

Supporting knowledge sharing in educational courses by means of CQA system Askalot

Successfulness of Community Question Answering (CQA) systems on the open web (e.g. Yahoo! Answers) motivated for their utilization in new contexts (e.g. education or enterprise) and environments (e.g. inside organizations). In spite of initial research how their specifics influence design of CQA systems, many additional problems have not been addressed so far. Especially a poor flexibility and scalability which hamper CQA essential features to be employed in various settings (e.g. in different educational organizations). In this paper, we provide design recommendations how to achieve flexible and scalable deployment by means of a case study on educational and organizational CQA system Askalot. Its universal and configurable features allow us to deploy it at several universities as well as in MOOC system edX.

Návrh a implementácia aplikácie pre tvorbu prezenčných listín pomocou mobilného zariadenia s technológiou NFC

Zuzana Vantová, Vladimír Gašpar

Fakulta elektrotechniky a informatiky
Technická univerzita v Košiciach
Letná 9, 041 20 Košice, Slovenská republika

`zuzana.vantova@student.tuke.sk, vladimir.gaspar@tuke.sk`

Abstrakt. Tento článok približuje problematiku tvorby prezenčných listín na Technickej univerzite v Košiciach. Taktiež približuje možnosti optimalizácie tohto procesu, vedúce k výslednému riešeniu. Riešenie je realizované v podobe mobilnej a webovej aplikácie s centrálnou databázou a REST API pre umožnenie komunikácie medzi komponentmi aplikácie, konkrétnejšie medzi centrálnou databázou a mobilnou aplikáciou. ISIC (International Student Identity Card) karta, ktorú musí vlastniť každý študent Technickej univerzity je v tomto prípade vhodným prostriedkom pre jednoznačné určenie prítomnosti študenta pomocou mobilného zariadenia s technológiou NFC. Prínosom riešenia je optimalizovaný proces tvorby prezenčných listín a zabezpečenie správnosti a úplnosti údajov. Na dosiahnuté výsledky je možné v budúcnosti nadviazať a pokračovať v rozširovaní aplikácie.

Typ príspevku: Aplikčný príspevok

Kľúčové slová: prezencie, aplikácia, NFC

1 Úvod

Na Technickej univerzite v Košiciach, ako aj na iných vysokých školách a univerzitách je evidencia prítomnosti študentov na cvičeniach, resp. aj prednáškach podmienkou udelenia zápočtu alebo skúšky. Je možné konštatovať, že vo väčšine prípadov sú prezenčné listiny vytvárané v papierovej podobe, čo znamená, že môže dôjsť k chýbam počas procesu ich vytvárania, prípadne je možné tento dokument ľahko stratiť. Hlavnou motiváciou pre aplikačný výstup, ktorý tento článok popisuje, je zjednodušenie a digitalizácia procesu vytvárania spomínaných prezenčných listín. Z technologického po-

hľadu je možné využívať študentské RFID karty, známe ako ISIC (International Student Identity Card) a na strane aplikácie komunikačnú technológiu NFC (Near field communication)¹²³.

Rozhodli sme sa pre vytvorenie prostredia, resp. aplikácie, ktorá zastreší jednak klientsku stranu (snímanie kariet a vytváranie pohľadu na prezencie študentov) ako aj administratívnu časť (správu zoznamov študentov, cvičení a prednášok). Integrovaným aspektom celého riešenia je databáza, ktorá poskytuje prístup k dátam mobilným klientom ako aj administrátorovi na webe prostredníctvom REST API služieb. Za zmienku stojí aj použitá technológia. Mobilná aplikácia bola vytvorená v prostredí Android Studio v jazyku Java⁴, zatiaľ čo web ako aj REST API v prostredí Visual Studio 2013 v jazyku C# použitím frameworku ASP.NET MVC 4⁵⁶.

2 Analýza problematiky

Ako sme už v úvode uviedli, proces vytvárania prezenčných listín je realizovaný podpisovaním sa študentov alebo kontrolou zo strany vyučujúceho. Existuje viacero nedostatkov, ktoré uvedené spôsoby vytvárajú a stručne sme ich spomenuli v úvode. Ako najvýhodnejšie riešenie sa javí možnosť využiť smartfón vyučujúceho a dáta synchronizovať s externou databázou. Takýmto spôsobom vieme zaručiť, že dáta budú aktuálne v každom čase a prezenčné listiny budú závislé primárne na ISIC kartách študentov. Môžeme tak konštatovať, že aktuálne spôsoby vytvárania prezenčných listín nie sú ideálne z pohľadu perzistencie, aktuálnosti a dostupnosti. Nami vytvorená aplikácia zlepši dostupnosť prezenčných listín, pretože budú jednak uložené centrálné v databáze ako aj v lokálnej databáze mobilného zariadenia. Aktuálnosť dát na oboch stranách zabezpečí synchronizácia údajov počas online režimu aplikácie. Nevyhnutnosťou je implementácia funkcie pre manuálny zápis študenta bez použitia NFC ako záložný spôsob pre prípady ak študent kartu zabudne, poškodí sa, resp. z iných dôvodov nie je možné využiť technológiu snímania kariet.

¹ Prelovský, Lukáš a iní: Čo je to vlastne NFC a aké má využitie. Online: <<http://www.lukasprelovsky.sk/co-je-to-vlastne-nfc-a-ake-ma-vyuzitie/>>

² NFC. Online: <<http://www.nearfieldcommunication.org>> 6.4.2016

³ Mój android: Začínáme s NFC: Čo je NFC a ako funguje. Online: <<https://www.mojandroid.sk/nfc-co-to-je-ako-funguje/>> 6.4.2016

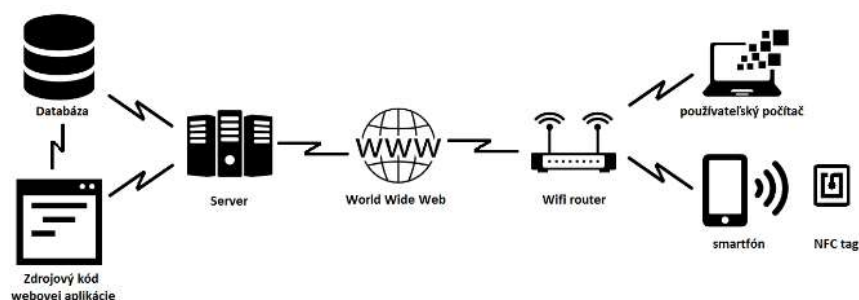
⁴ Android Developers: Android, the world's most popular mobile platform. Online: <<http://developer.android.com/about/android.html>> 2.4.2016 (v anglickom jazyku)

⁵ MVC architektúra. Online: <<http://www.itnetwork.cz/navrhove-vzory/mvc-architektura-navrhovy-vzor>> 5.4.2016

⁶ Coplien, James O., Reenskaug Trygve: The DCI Architecture: A New Vision of Object-Oriented Programming, 20.3.2009, Online <http://www.artima.com/articles/dci_vision.html> 28.6.2016 (v anglickom jazyku)

3 Implementácia aplikácie

V prvom rade bolo potrebné navrhnuť konceptuálnu architektúru riešenia (Obr. 1) a získať tak jednoduchý pohľad na riešenie problematiky. Na základe tohto návrhu sme postupne vytvárali aplikačné prostredie pre jednotlivé využitia aplikácie. Najprv bolo potrebné navrhnuť a vytvoriť databázu, ktorá bude zjednocujúcim prvkom jednotlivých aplikácií. Táto databáza v prvotnom návrhu disponovala štyrmi tabuľkami, ktoré poskytovali základné informácie o cvičeniach, prednáškach, študentoch a priradení študentov na prednášky a cvičenia.



Obr. 1 Architektúra riešenia

Po dotvorení finálnej verzie databázy bola vytvorená webová aplikácia pre správu študentov, prednášok, cvičení a ich priradzovanie na cvičenia a prednášky. Na základe testovania bol doladený aj dátový model a doplnená celková idea o aplikácii.

Pre webovú aplikáciu sme zvolili formu „*long scroll*“, nakoľko v súčasnosti patrí medzi často používané trendy. V rámci stránky sa nachádza viacero sekcií zobrazujúcich údaje o prítomnostiach študentov na prednáškach a cvičeniach, ako aj priradenie jednotlivých cvičení k študentom. Prítomnosť študentov na prednáškach a cvičeniach je menená prostredníctvom stavového tlačidla (prítomný-nepítomný-bez záznamu).

Technicky je interaktivita webovej aplikácie zabezpečená jQuery a Ajax⁷⁸ funkciami napríklad pre zobrazenie upozornení alebo informovaní pri ukladaní dát a ich zápise do databázy, pri prepínaní zoznamov študentov medzi jednotlivými cvičeniami a pod. Zaradenie študentov na cvičenia sa vykonáva zaškrtnutím *radiobutton*-ov. V rámci tejto sekcie je umožnené pridávať nových študentov ako aj nové cvičenia.

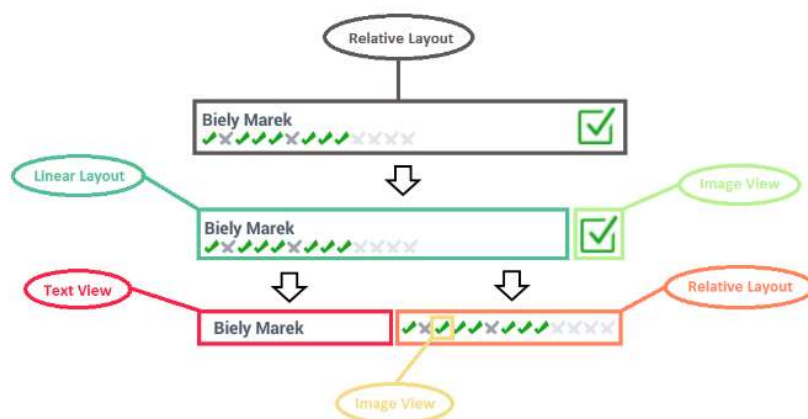
⁷ Taft, Darryl K.: jQuery Eases JavaScript, AJAX Development, 30.8.2006, Online: <<http://www.eweek.com/c/a/Application-Development/jquery-Eases-JavaScript-AJAX-Development>> 28.6.2016 (v anglickom jazyku)

⁸ MDN Mozilla Developer Network: Ajax. Online: <<https://developer.mozilla.org/en-US/docs/AJAX>> 5.4.2016 (v anglickom jazyku)

Funkcie v REST API sú vytvorené ako samostatný *controller* webovej aplikácie. Slúžia pre zabezpečenie komunikácie medzi mobilnou aplikáciou a databázou. Žiadosti, ako aj odpovede, sú odosielané vo formáte JSON⁹.

Mobilnú aplikáciu tvoria 4 obrazovky - *layouty*. Umiestnené sú vo *FrameLayout*-e a svojou štruktúrou sú podobné webovej aplikácii. Vynechali sme zatriedenie študentov, pretože z pohľadu používateľského zážitku (UX) na mobile by bolo nevyhovujúce.

Prvý *layout* (Úvod) obsahuje informácie o priemernej prítomnosti študentov na prednáškach a cvičeniach. Tieto informácie používateľ môže zobrazit' vo forme čiarového grafu kopírujúceho priebeh semestra. Ďalšími *layout*-mi sú Prednášky a Cvičenia a zobrazujú prítomnosti študentov počas semestra. Zoznam študentov je tvorený hierarchiou objektov *Relative Layout*, *Linear Layout* a *Image View*. Hierarchia je zobrazená na nasledujúcom obrázku.



Obr. 2 Štruktúra informácií o študentovi

Stavy prítomnosti sa menia podobne ako vo webovej aplikácii, teda klikaním na stavové tlačidlá prítomnosti. Zmena stavu pre vybraný týždeň sa zobrazí pod menom študenta, ako aj v pravej časti obrazovky. Zápis študentov je možné vykonávať aj prostredníctvom NFC technológie, priložením ISIC karty k NFC čítačke smartfónu a následne sa daný študent zapíše do prezencie. Úspešný zápis je znázornený správou, *toast*-om, v dolnej časti obrazovky, ktorá obsahuje meno študenta a UID ISIC karty.

4 Záver

Článok popisuje návrh a implementáciu aplikácie pre vytváranie a správu prezenčných listín. Na strane mobilného klienta bola použitá technológia NFC pre snímanie RFID

⁹ JSON: The Fat/Free Alternative to XML, Online: <<http://www.json.org/xml.html>> 28.6.2016 (v anglickom jazyku)

kariet študentov a na pozadí bola vytvorená webová aplikácia a databáza pre udržiavanie takto zozbieraných údajov. Predpokladáme, že sa aplikácia bude využívať v praxi už nasledujúci semester po jej otestovaní a odstránení chýb. Testovanie bude súčasťou tohto využitia. Plánujeme zaviesť uvedený spôsob kontroly prítomností minimálne na jednom cvičení, pričom efektívnosť a správnosť využitia riešenia bude porovnaná s klasickým prístupom kontroly pred koncom semestra dotazníkovou formou, ako pre študentov, tak aj pre vyučujúcich.

Vytvorenie tejto aplikácie bolo cieľom diplomovej práce.

Podakovanie: Táto publikácia vznikla vďaka podpore v rámci operačného programu Výskum a vývoj pre projekt "Centrum informačných a komunikačných technológií pre znalostné systémy" (kód ITMS:26220120020), spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja (50%). Publikácia bola zároveň podporená projektom KEGA č. 014TUKE-4/2015 "Digitalizácia, virtualizácia a testovanie malého prúdového motora pomocou stendov pre potreby modernej aplikovanej výuky" (50%).

Literatúra

1. Prelovský, Lukáš a iní: Čo je to vlastne NFC a aké má využitie. Online: <<http://www.lukasprelovsky.sk/co-je-to-vlastne-nfc-a-ake-ma-vyuzitie/>>
2. NFC. Online: <<http://www.nearfieldcommunication.org>> 6.4.2016
3. Môj android: Začínáme s NFC: Čo je NFC a ako funguje. Online: <<https://www.mojandroid.sk/nfc-co-to-je-ako-funguje/>> 6.4.2016
4. Android Developers: Android, the world's most popular mobile platform. Online: <<http://developer.android.com/about/android.html>> 2.4.2016 (v anglickom jazyku)
5. MVC architektúra. Online: <<http://www.itnetwork.cz/navrhove-vzory/mvc-architektura-navrhovy-vzor>> 5.4.2016
6. Coplien, James O., Reenskaug Trygve: The DCI Architecture: A New Vision of Object-Oriented Programming, 20.3.2009, Online <http://www.artima.com/articles/dci_vision.html> 28.6.2016 (v anglickom jazyku)
7. Taft, Darryl K.: jQuery Eases JavaScript, AJAX Development, 30.8.2006, Online: <<http://www.eweek.com/c/a/Application-Development/jquery-eases-javascript-ajax-development>> 28.6.2016 (v anglickom jazyku)
8. MDN Mozilla Developer Network: Ajax. Online: <<https://developer.mozilla.org/en-US/docs/AJAX>> 5.4.2016 (v anglickom jazyku)
9. JSON: The Fat/Free Alternative to XML, Online: <<http://www.json.org/xml.html>> 28.6.2016 (v anglickom jazyku)
10. GS1 Slovakia: RFID na Slovensku. 10s. Online: <http://www.gs1sk.org/down/RFID_na_Slovensku.pdf> 2.4.2016
11. Kodys Slovensko: RFID. Online: <<http://www.kodys.sk/stranka/rfid>> 2.4.2016
12. Howard, M.; LeBlanc, D.: Bezpečný kód; Computer Press, a.s., 2008. 888 s. ISBN 978-80-251-2050-7.
13. Galloway, J.; Wilson, B.; Allen, K. S.; Matson, D.: Professional ASP.NET MVC 5; Wrox Publishing, 2014. 624 s. ISBN 978-1-118-79475-3.
14. Goldsteinová, A.; Lazaris, L.; Weylová, E.: HTML5 a CSS3 pro webové designéry; Zoner Press, 2011. 288 s. ISBN 978-80-7413-166-0.
15. Bruckner, T.; Voříšek, J.; Buchalceová, A. a kol: Tvorba informačních systémů; Grada Publishnig, a.s., 2012. 360 s. ISBN 978-80-247-4153-6.

Annotation:

Design and implementation of a application for creating presence lists using a smartphone with NFC

This paper presents a problem of creating presence lists at the Technical university in Košice. It also describes possibilities of optimizing this process, which leads to the final solution. The solution is realized in form of a mobile and web application with a central database and REST API for communication between components of the application, particularly between the central database and the mobile application. ISIC card, which every student of Technical University must possess, is in this case a suitable medium for clear identification of student's presence using a smartphone with the NFC technology implemented. The asset of this solution is the optimized process of creating presences and enhancement of the correctness and completeness of data. It is possible to continue in extending the application in the future.

Využití cloudu pro dolování asociačních pravidel z velkých dat přes webové rozhraní

Václav Zeman¹, Stanislav Vojír¹, Jaroslav Kuchař^{1,2}, Tomáš Kliegr¹

¹ Katedra informačního a znalostního inženýrství
Fakulta informatiky a statistiky
Vysoká škola ekonomická v Praze
Nám. W. Churchilla 4, 130 67 Praha 3, Česká republika
² Fakulta informačních technologií
České vysoké učení technické v Praze
Thákurova 9, 160 00 Praha 6, Česká republika

{vaclav.zeman, stanislav.vojir, jaroslav.kuchar,
tomas.kliegr}@vse.cz

Abstrakt. Webová aplikace EasyMiner je akademický nástroj pro získávání znalostí z malých a středně velkých dat ve formě asociačních pravidel. Nová verze tohoto systému využívá prostředí Apache Hadoop a Apache Spark pro zpracování velkých datových zdrojů na výpočetním clusteru MetaCentra sdružení CESNET. Aplikace se skládá z několika mikro služeb, které se starají o nahrávání velkých dat do distribuovaného úložiště HDFS, transformaci dat v clusteru do normalizované formy a dolování znalostí z datasetů v podobě asociačních pravidel s využitím výpočetních prostředků clusteru pomocí nástroje Apache Spark. S těmito mikro službami se dá komunikovat prostřednictvím RESTového rozhraní a jako celek tvoří data miningový software fungující jako webová služba - SaaS.

Typ příspěvku: Aplikací příspěvek

Klíčová slova: data mining, dolování asociačních pravidel, big data, hadoop

1 Úvod

Akademický nástroj EasyMiner¹ je webová služba se zaměřením na dolování asociačních pravidel z databází [2]. Aplikace poskytuje grafické uživatelské rozhraní a je schopna vykonat všechny nutné operace pro získávání znalostí z dat od nahrávání datasetů přes předzpracování až po samotné dolování a interpretaci výsledků. Nová verze tohoto nástroje dokáže zpracovat i velká data díky nasazení do prostředí Apache Hadoop a Apache Spark a lze ji použít pro akademické účely zcela zdarma s využitím výpočetního clusteru na půdě MetaCentra² sdružení CESNET (až 24 uzlů x 16 jader x

¹ <http://www.easyminer.eu/>

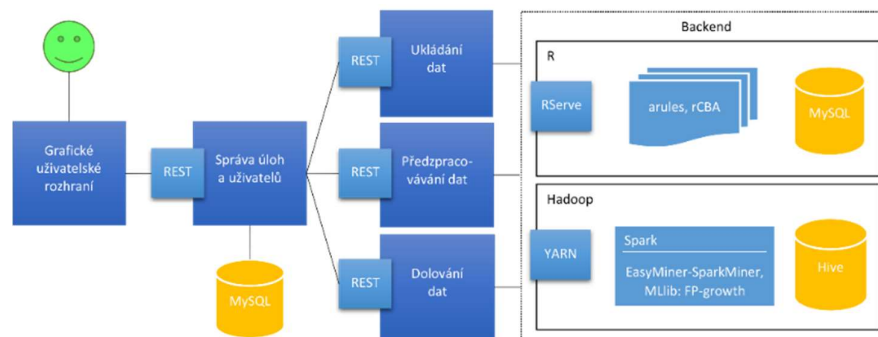
² <https://wiki.metacentrum.cz/wiki/Hadoop>

2 hyperthreading s více než 3TB RAM). Mezi nejdůležitější operace, které lze v systému EasyMiner vykonávat, patří:

- Proudové nahrávání datových zdrojů do datového úložiště
- Předzpracování dat pro účely rychlejšího dolování znalostí
- Dolování asociačních pravidel dle uživatelských požadavků
- Tvorba klasifikačních modelů ze získaných pravidel
- Manipulace s množinou získaných pravidel

Nástroj lze tedy v kontextu cloudových služeb zařadit do kategorie MLaaS (Machine Learning as a Service) a lze jej rovněž použít jako alternativu ke komerčním produktům, jako je např. BigML.com či Microsoft Azure ML, která je více orientovaná na pravidla.

Aplikace jako taková se skládá z několika mikro služeb, které spolu navzájem komunikují skrze RESTová API (viz Obr. 1). Většina těchto mikro služeb pracuje ve dvou různých režimech *limited* a *unlimited*. Režim *limited* slouží pro správu malých a středně velký datasetů, přičemž je využito databáze MySQL jako primárního datového skladu a prostředí R pro dolovací účely. V režimu *unlimited* komunikují služby převážně se systémem Apache Hadoop, který je využíván hlavně k ukládání velkých dat pomocí nástroje Apache Hive a pro distribuované hledání asociačních pravidel postavené na frameworku Apache Spark.



Obr. 1 Architektura systému EasyMiner

2 Podpora velkých dat

Primární backendové řešení je postaveno na knihovně *arules* z prostředí R [3, 4]. Toto řešení je bohužel špatně škálovatelné a není vhodné pro použití na velkých datech; proto byla naimplementována další backendová vrstva, která se specializuje výhradně na dolování pravidel z velkých dat.

Pro ukládání a předzpracování dat je využíváno rozhraní Apache Hive. Samotné dolování probíhá jako Spark úloha. Veškeré distribuované úlohy jsou spravovány systé-

mem YARN. Tato architektura je založena na dávkovém vykonávání jednotlivých distribuovaných operací, tudíž se příliš nehodí pro menší datasety, které mohou být zpracovány na jednom stroji mnohem rychleji díky in-memory real-time přístupu. Silné a slabé stránky jednotlivých backendových řešení lze vyčíst ze srovnávací tabulky 1.

Tab. 1 Srovnání dvou backendových vrstev v systému EasyMiner

| Vlastnosti | Backend | |
|--|-------------------------------------|--------------------------------------|
| | Limited | Unlimited |
| Prostředí | R | Apache Hadoop, Apache Spark |
| Úložiště | MySQL | HDFS + Apache Hive |
| Forma uložených dat | řádkově orientované tabulky v RDBMS | sloupcově orientované tabulky v HDFS |
| Čas vykonávání úlohy | sekundy | desítky sekund až desítky minut |
| Velikost dat | do 100MB | více než stovky MB |
| Škálovatelnost úlohy | ne | ano (počet uzlů x počet jader) |
| Paralelní úlohy | ano (závisí na počtu vláken) | ano (závisí na YARN plánovači) |
| Přístup dolování | in-memory | distribuovaně in-memory |
| Algoritmus pro dolování asociačních pravidel | apriori (knihovna arules) | FP-growth (knihovna MLlib) |
| Algoritmus pro tvorbu klasifikačních modelů | CBA | CBA |

3 Zpracování a dolování dat

Data lze do příslušného úložiště nahrávat skrze datovou službu, která umožňuje proudové ukládání dat buď do MySQL pro *limited* režim, nebo do HDFS pro *unlimited* režim. Velikost nahrávaného datasetu není v hadoop prostředí nijakým způsobem omezena. Pro snazší a rychlejší zpracování dat jsou data v HDFS ukládána do sloupcově orientované podoby [1]. Díky takovéto reprezentaci lze provádět agregační a joinovací funkce napříč všemi sloupci jednou MapReduce úlohou (viz srovnání v Ttab. 2).

V procesu předzpracování dat dochází k mapování všech hodnot na číselné indexy, čímž dochází k mírné kompresi dat a samotné dolovací algoritmy mohou pracovat pouze s jedním datovým typem bez nutnosti řešení kódování textu.

Hledání asociačních pravidel z menších dat je primárně prováděno v prostředí R. Toto řešení je velmi rychlé, avšak vyžaduje na vstupu kompletní databázi transakcí uloženou v paměti. V distribuovaném prostředí se pro dolování asociačních pravidel využívá knihovna Spark MLlib, konkrétně algoritmus FP-growth [5].

Tvorba klasifikačního modelu z nalezených pravidel je součástí dolovacího Spark programu, který implementuje algoritmus CBA [6]. Tento algoritmus prořezává nalezená pravidla, za kterými je připojeno tzv. výchozí pravidlo. Z tohoto výstupu se poté dá vytvořit klasifikátor pro zvolenou cílovou proměnnou.

Tab. 2 Srovnání řádkově a sloupcově orientovaných Hive tabulek při použití v systému EasyMiner. Proměnná N vyjadřuje počet sloupců v tabulce.

| Operace | Počet MapReduce úloh | |
|-------------------------------|-----------------------------|-------------------------------|
| | Řádkově orientovaná tabulka | Sloupcově orientovaná tabulka |
| Ukládání datového zdroje | $2N + 1$ | 2 |
| Čtení agregovaného histogramu | 1 | 1 |
| Tvorba datasetu | 1 | 0 |
| Předzpracování sloupců | $N/5 + 3$ | 3 |

4 Závěr

Ačkoliv je cloudová verze nástroje EasyMiner stále ve fázi vývoje, je možné ji používat pro testovací a akademické účely bez jakýchkoliv omezení na MetaCloudu sdružení CESNET. Aplikace je schopna stabilně nahrávat data a hledat v nich asociační pravidla, ze kterých lze sestavovat klasifikační modely. Budoucí vývoj je v současném stavu zaměřen na zrychlení dolovacích algoritmů, implementaci diskretizačních algoritmů a nasazení nástroje pro hledání anomálií. V plánu je také podpora dolování pravidel z RDF.

Poděkování: Tato práce vznikla za podpory Vysoké školy ekonomické v Praze pod grantem IGA 29/2016 a Fondu rozvoje CESNET pod grantem č. 540/2014.

Literatura

1. Abadi, D. J., Madden, S. R., Hachem, N.: Column-stores vs. row-stores: How different are they really?. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD '08, (2008), pp. 967-980.
2. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, SIGMOD '93, (1993), pp. 207-216.
3. Hahsler, M., Buchta, C., Grün, B., Hornik, K. arules: Mining Association Rules and Frequent Itemsets, (2010). URL <http://cran.r-project.org/package=arules>. R package.
4. Kuchar, J., Kliegr, T., Vojír, S., Zeman, V.: EasyMiner/R Preview: Towards a Web Interface for Association Rule Learning and Classification in R. In Rule Challenge and Doctoral Consortium @ RuleML 2015, (2015)
5. Li, H., Wang, Y., Zhang, D., Zhang, M., Chang, E. Y: PFP: parallel fp-growth for query recommendation. In Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys '08, (2008), pp. 107-114.
6. Liu, B., Hsu, W. Ma, Y.: Integrating classification and association rule mining. In Proceedings of the fourth international conference on knowledge discovery and data mining, KDD '98, AAAI Press, (1998), pp. 443-447.

Annotation:

Association Rules Mining for Big Data in Cloud

EasyMiner is a web service for association rules mining. A new version of this tool uses Apache Hadoop and Apache Spark for big data processing in the MetaCloud of the CESNET association. The application consists of several services for dataset uploading into HDFS, preprocessing, association rules discovery and classification based on associations. All services communicate with each other through REST APIs and form a complex software working as a service in the cloud.

Smerovanie dizertačných projektov

Web-Human Interaction Based on Ontology Query

Jana Ahmad, Petr Křemen

Faculty of Electrical Engineering
Czech Technical University in Prague
Žitkova 1903/4166 36 Prague 6, Czech Republic

{jana.ahmad, petr.kremen}@fel.cvut.cz

Abstract. In recent years, human-ontology interaction becomes an increasingly important subject for computational and information systems developers. Human information consumers and web agents need to use and query ontologies using their web sites and web applications, thus the need for developing tool supporting ontological engineering and querying tools arises. In this paper we discuss the potential of web based human-readable ontology queries. In large taxonomies such Aviation Safety (AS) domain, it is important to allow easier navigation within the ontology. Thus we introduce an extension to the OntoQuery tool for more practical visualization of query results and easy OWL vocabulary dissemination to the community.

Contribution type: PhD Symposium

Keywords: ontology, query, aviation safety

1 Introduction

In last years, ontology has been applied in a large number of areas in computer science. It is also used to refer to specific material domains (e.g., medicine, biology, aviation safety, etc.), resulting in domain ontologies. During ontology design and exploitation users need exploring ontology structure. Web-based tools are well suitable for this. Due to complex structure of ontologies, there is a need to pay attention to the development of human-ontology-interaction, and enhancing the querying ontology tools.

In this paper we discuss how applying our OntoQuery extension to a specific domain ontology (aviation safety domain) could help this domain's users and experts to explain, evaluate and exploit vocabularies in this domain.

Section 2 presents motivations and related work. Section 3 discusses our extension to OntoQuery tool. In subsection 3.1 the use case used in our research is presented. Finally, Section 4 concludes this work.

2 Motivation and Related Work

There are several different techniques and tools supporting user interaction with ontologies. These tools help user and ontology experts to query and explore vocabularies and concepts. However most ontological engineering tools suffer from different problems w.r.t interaction with human user, as discussed next.

Protégé-OWL [5] is a knowledge based ontology editor providing graphical user interface. It provides flexibility for meta-modeling and enables the construction of domain ontologies. But some studies found that visualization options offered by Protégé are too complex. Also, many users have difficulties with description logic (even though Manchester syntax is used).

OntoQuery tool [2] introduces the OntoQuery web-based query utility. The interface of OntoQuery tool provides syntax highlighting similar to that provided by the Protégé DL query tool. However, unlike Protégé, OntoQuery highlighting distinguishes between classes and properties. As the user types, the system pops up a box with suggestions appropriate to the syntactic position within the query. The queries will return all descendants (not just direct subclasses) matching the logical definition expressed in Manchester syntax.

OWLGrEd [4] the OWLGrEd Ontology Visualizer is an online tool for visualizing OWL ontologies using a compact UML-based notation.

In our previous work, we based our aviation safety vocabulary explorer [9], which aims to visualize and explore the concepts (classes and relations) of aviation safety domain. It also helps aviation safety websites users to clearly understand the aviation safety vocabularies, in order to make their safety reports more efficient. We realized the importance of making navigation within large taxonomies easier for user. Thus we added extension to OntoQuery tool, which aims to categorize aviation safety domain into categories according to most general concepts, in order to facilitate the navigation within the aviation domain for domain's users and experts.

3 Extension to OntoQuery Tool

Our plugin extension to OntoQuery tool [2] selects intentional classes from our domains and corresponding conceptual (presented in section 3.1) using Protégé DL query and named categories. It categorizes our ontology by adding `isSubcategoryOf` annotation property to each concept regarding to its category. When the user types his query in the client-side JavaScript input box, this query is sent to the server, which checks the syntax of the query, the translation of labels to IDs and the parsing of the query to OWL Manchester syntax are performed on the server. Then the query is executed to categorize ontology vocabularies according to the categories, that we selected by querying annotation property. However, this extension is only beneficial for large taxonomies (e.g., Aviation Safety ontology). Thus, simple domain-specific categorization of ontology terms allows easier navigation within the ontology.

It is important and helpful for aviation safety agents to get details and good explanation about their input during searching in online safety websites. Thus, as a concrete

example of our work (see figure 1), when user uses aviation vocabulary explorer interface to search for some vocabulary terms (e.g., Airborne Object), it is very beneficial and helpful not only for users, but also for aviation safety systems and experts to give to user more information and explanation about his input by using ontology categorizations (e.g., Airborne Object is related to Physical Object category, which seems understandable for human recognition).

Aviation Vocabulary Explorer

Collision and has_participant some Airborne_object

Quick Tips Examples Recent Queries Tutorial Results 2

Filter Results

| Vocabulary | Title | Description | Categories |
|-------------------|-----------------|-------------|-----------------|
| Aviation Ontology | Airborne object | | Physical object |
| Aviation Ontology | Collision | | Event |

Fig 1. Aviation safety explorer categorization

3.1 3.1 Use case

In this paper we consider the following domains and corresponding conceptual models:

- A Conceptual Model representing the domain of aviation safety (1737 classes). It defines general well understood concepts in Aviation domain such as Aircraft, Flight, Agents and etc [10].
- A Conceptual Model that describes Eccairs taxonomies ontology (4067 classes). It aims to Improve air safety by bringing together the knowledge derived from the collection of incompatible occurrence reporting systems from various (member) States [8].
- Unified Foundational Ontology (UFO), which is a top-level ontology. UFO is an ontology for specifications of domain ontologies and languages. It is divided into three layers: Object and Trope model part (UFO-A) [1], event model part (UFO-B) [6] and service model part (UFO-C) [7].

We select our categories w.r.t the most general and understandable taxonomies in models that we mentioned above, we selected the most twenty (20) general concepts as categories. For example, we select: physical object and data categories which relate to aviation safety ontology. Event and trope categories relate to UFO concepts.

4 Conclusion

In this paper we discussed how developing query based on ontology (by extending On-toQuery tool), especially in very large taxonomies (e.g., aviation safety) could help domain's users and experts to navigate within ontology in very easy way.

Acknowledgements: This work was partially supported by grants No. TA04030465 Research and development of progressive methods for measuring aviation organizations safety performance of the Technology Agency of the Czech Republic, No. SGS16/229/OHK3/3T/13 Supporting ontological data quality in information systems of the Czech Technical University in Prague, and No. GA 16-09713S Efficient Exploration of Linked Data Cloud of the Grant Agency of the Czech Republic.

References

1. Guizzardi, G. Ontological Foundations for Structural Conceptual Models, PhD Thesis (CUM LAUDE), University of Twente, the Netherlands. Published as the book Ontological Foundations for Structural Conceptual Models, Telematica Instituut Fundamental Research Series No. 15, ISBN 90-75176-81-3 ISSN 1388-1795; No. 015; CTIT PhD-thesis, ISSN 1381-3617; No. 05-74.
2. Ilinca, T., Christoph. S. OntoQuery: Easy-to-use web-based OWL querying in Bioinformatics September 2013.
3. The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology Nat. Genet., 25, 25–9.
4. <http://owlgred.lumii.lv/>. Cited 11-06-2016.
5. Horridge. A Practical Guide to Building OWL Ontologies Using Protege 4 and CO-ODE Tools Edition 1.3 Available at http://mowl-power.cs.man.ac.uk/protegeowltutorial/resources/ProtegeOWLTutorialP4_v1_3.pdf.
6. G Guizzardi and G Wagner. Towards ontological foundations for the conceptual modeling of events. In 32th International Conference, ER 2013, Hong-Kong, China, November 11-13, 2013. Proceedings, 2013.
7. Julio Cesar Nardi, Ricardo de Almeida Falbo, Joao Paulo A. Almeida, Giancarlo Guizzardi, Luis Ferreira Pires, Marten J. van Sinderen, and Nicola Guarino. Towards a Commitment Based Reference Ontology for Services. In 2013 17th IEEE International Enterprise Distributed Object Computing Conference.
8. <http://eccairsportal.jrc.ec.europa.eu>. Cited 6.13.2016.
9. <https://www.inbas.cz/aviation-vocabulary-explorer>. Cited 6.20.2016.
10. <http://www.inbas.cz/ontologie>. Cited 6.20.2016.

miXGENE: an Effective Public Tool for Integrative Analysis of High-throughput Omics Data

Michael Anděl, Pavel Strnad, Jiří Kléma

Department of Computer Science, Faculty of Electrical Engineering
Czech Technical University in Prague
Technická 2, 166 27 Praha, Česká republika

{andelmi2, strnapal, klema}@fel.cvut.cz

Abstract. Molecular biology is a domain endowed by a good amount of data and well-formalized knowledge. Based on measured data and domain knowledge, an intelligent integrative analysis is capable of extracting new and more specific knowledge, which may help to comprehension of e.g. disease mechanisms. We have proposed miXGENE, a web service for integrating and analyzing high-throughput omics data, namely from the microarray-based expression or methylation measurements, together with formal biological knowledge, such as gene ontologies and curated or predicted omics interactions. The tool enables building the most employed analytical workflows for processing user-data or the data from public databases. Processing of the data is followed by their integrative statistical or machine-learning based analysis, and completed with the presentation of results in the expert-comprehensible terms. We propose an innovation of the tool which profits from the infrastructure of the Czech National Grid, CESNET – MetaCentrum, which facilitates the most computationally demanding sections.

Contribution type: PhD Symposium

Keywords: omics data, web service, machine learning

1 Introduction

One of the key paradigm in data-mining research and practice is integration of *formal knowledge* related to the investigated domain [1] into the analytical process. This knowledge incorporation is expected to help in making more precise models, interpreting the models and discovering new knowledge. In the other words, when having a loose notion what we are searching for, we can adjust (bias) our algorithm towards this particular knowledge. The resulting model would be more specific to the researched domain, and thus interpretable in the predefined terms and potentially unveiling a new knowledge yet more specific knowledge related to the researched problem. Last but not least, the more accurate models are expected from this paradigm, as the knowledge by

restricting the space of all conceivable hypothesis prevent overfitting and induces more generalization.

One of the domains with a lot of generated data and well-formalized *domain knowledge* is molecular biology. Thanks to current technological progress in microarrays and next-generation sequencing, we are being flooded with high-throughput measurements related to the *genome*, *transcriptome* and *proteome*. A neologism describing the measurements from all these biological sources is *omics* data. The genome is the set of individual's genetic equipment encoded in its DNA. The particular amount of genes *transcript*, which is further translated into *protein* is the fundamental process, called *gene expression* (GE), of migrating the biological information from DNA to the visible signs called *phenotype*. The *knowledge* linking all these processes and components is in the form of predicted or validated omics interactions, gene ontologies or curated canonical pathways and gene sets. By intelligent integration of these data types and by incorporating related knowledge we can extract a valuable nuggets which may help to comprehension of e.g. disease mechanisms or ordinating a *personalized* treatment.

We have proposed miXGENE [2], a web service for integrating and analyzing high-throughput omics data, namely the *microarray*-based measurements of mRNA and microRNA expressions, and the *methylation* assays, together with formal biological knowledge mentioned above. The expression data sets are possible to be uploaded by the user, or to be fetched from the public database NCBI GEO (National Center for Biotechnology Information – Gene expression Omnibus) [3]. The knowledge is internally represented as graphs of omics interactions or sets of the omics units, and may be uploaded by the user in a predefined canonical format. Otherwise, the user can choose default system-based knowledge sources, which had originated from the already curated sources, namely from the gene ontologies (GO) [4], KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways and other curated gene sets from the Molecular Signature Database (MSigDB) [5]. The tool is conceived as a *workflow management system*, which enables building the most employed analytical pipelines. Processing of the data is followed by their integrative statistical or machine-learning based analysis. The workflow is completed with the presentation of results in the expert-comprehensible terms and visualization.

To make the tool more effective, namely for large-scale bioinformatics experiments, we have migrated the most demanding segments of the computation to the infrastructure of Czech National Grid. The tool is freely available at mixgene.felk.cvut.cz/.

2 Related Work

The workflow management systems (WMS), which miXGENE is an instance of, are growing area of research [6]. The main purpose of those systems is: (i) to make the computational biology accessible for those researchers who are instructed informaticians yet not programmers, (ii) to enable tracking of experimental history and offer a tool for testing different settings, and (iii) the possibility to exchange the scientific workflows. There are many general tools designed to represent bioinformatic or data-

analytic workflows; e.g., Taverna [7] or Galaxy [8]. miXGENE has been implemented as a specialized bioinformatics WMS. To facilitate the most computationally demanding workflows, we migrate the critical parts to the infrastructure of the Czech National Grid, CESNET – MetaCentrum [9].

There are several bioinformatics tools, residing in the CESNET structures and integrating number of tools into a computational pipeline [10], [11]. However, miXGENE is a tool based on independent server, which operatively migrates the most demanding computational segments into the CESNET infrastructure.

3 System description

3.1 Basic Architecture

The tool can be split into three parts: 1) GUI (task definition, presentation of results), 2) workflow management (task decomposition and its global planning in terms of the individual plugins) and 3) computational plugins (implementation of the individual analytical methods such as data normalization, feature extraction, learning of classifiers, etc.). Web interface and storage management are implemented in the web application framework Django, the workflow management is implemented in JavaScript and the computational plugins are mainly implemented in Python.

With miXGENE, all experiments are built from components called blocks using interactive workspace. miXGENE defines two types of blocks: the proper-blocks and *meta-blocks*. Former represents particular *atomic* tasks.

Each block represents one meaningful step in the experiment such as: 1) providing data source, i.e., user-uploaded or fetched dataset (set of measured expressions of genes) and/or a source of formal knowledge (interaction graph, curated gene sets or pathways); 2) preprocessing, analysis and creating the model itself; 3) presentation of the model and its results. The execution order is inferred from the data flow defined by binding the corresponding output and input ports of the consecutive blocks.

The meta-blocks serve as containers of blocks or other meta-blocks. They generate their own scope of possible input variables. Actually, for a single sequence of blocks, the metablocks create alternative scenarios over multiple inputs, variables or parameterizations. The main use of the meta-blocks is custom iteration over an user-defined collection of: a) data sets or knowledge sources, b) subsets of a data set or c) different analytical methods and their configurations. Alternating these variables contributes respectively for: (i) validating a method or an analytical workflow over as much data inputs as possible, (ii) validating the method for a particular dataset (i.e. cross-validation) and (iii) assessing the reliability of a knowledge source, such as putative pathways or predicted omics interactions, which the user had previously acquired.

3.2 Migration of the computationally critical segments

Formerly, a single experiment was executed serially. It was packed as a whole and sent to the miXGENE application server. In this innovation, we decompose the workflow into smaller, mutually independent tasks, pack them and send them to the grid.

Particularly, the *cross-validation* meta-block consists of semantically independent sub-workflows (folds) of a common pattern. In our implementation, we pack the instances related to the sequence (sub-workflow) inside the cross-validation scope. The packed instances are then asynchronously sent to the grid nodes, where they are executed. The server is receiving the finished tasks and integrates them into a result container. This approach fits the *map-reduce* paradigm.

4 Conclusion

We propose an innovation of our workflow management system miXGENE. The system serves for easy-to-use construction of scientific pipelines for bioinformatical use. The innovation lies in the effective migration of the most time-consuming segments of workflows to the infrastructures of Czech National Grid.

Acknowledgment: The system miXGENE has been developed with a support of grant NT14539 of the Ministry of Health of the Czech Republic. The innovation of the system is granted by the CESNET Development Fund. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

References

1. Sinha, A., Huimin, Z.: Incorporating domain knowledge into data mining classifiers: An application in indirect lending. *Decis Support Syst*, (2007), vol. 46, no. 1, pp. 187-299.
2. Holec, M., Gologuzov, V., Klema J.: miXGENE tool for learning from heterogeneous gene expression data using prior knowledge. In: *Proc. of 2014 IEEE 27th International Symposium on Computer-Based Medical Systems*, IEEE Press, (2014), pp. 247-250.
3. Barret, et al.: NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Research*, (2013), vol. 41, no. D1, pp. D991-D995.
4. Ashburner, M., Ball, C. A., Blake J. A., et al.: Gene Ontology: tool for the unification of biology. *Nature Genetics*, (2000), vol. 25, no. 1, pp. 25-29.
5. Liberzon, et al.: Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, (2011), vol. 27., no 12, pp. 1739-1740.
6. Barker, A., Hemert, J. V.: "Scientific workflow: a survey and research directions. In *Proc. Of Parallel Processing and Applied Mathematics.*, Springer Berlin, (2007), pp. 746-753.
7. Wolstencroft, K., et al.: The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acid Research*, (2013), vol. 41. pp. W557-W561.

8. Goecks, J., et al.: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, vol. 11, no. 8., pp. R86.
9. Šustr, et al.: Metacentrum, the Czech Virtualized NGI. In *Proc. Of EGEE Technical Forum*, (2009).
10. Michalovová, et al.: Fully automated pipeline for detection of sex linked genes using RNA-Seq data. *BMC Bioinformatics*, (2015), vol. 16, no. 1.
11. Schmickl, et al.: Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). *Molecular Ecology Resources*, (2015).

Využívanie hierarchie vetných vzorov pre automatizovanú tvorbu otázok

Miroslav Blšták, Viera Rozinajová

Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava, Slovenská republika

{miroslav.blstak, viera.rozinajova}@stuba.sk

Abstrakt. V tomto príspevku predstavujeme metódu na automatizovanú tvorbu faktických otázok z textu, ktorá využíva informácie o štruktúre vety a sémantické informácie o slovách. Na základe štruktúry vety vytvárame vetné vzory pomocou ktorých transformujeme deklaratívne vety na otázky rôznych typov. Sémantické informácie o slovách zlepšujú kvalitu vygenerovaných otázok. Významným zlepšením oproti bežným prístupom je uchovávanie vzorov v hierarchii, čo umožňuje spravovať vzory efektívnejšie a generovať otázky rôznej úrovne abstrakcie. Na záver navrhujeme využiť strojové učenie na vytváranie a odvodzovanie nových vzorov využívajúcich syntaktickú štruktúru viet a sémantické kategórie slov.

Typ príspevku: Doktorandské sympóziu

Kľúčové slová: automatizovaná tvorba otázok, vetné vzory, spracovanie textu

1 Úvod

Kvalita a množstvo vzdelávacích materiálov dostupných online neustále rastie. Vzdelávanie pomocou týchto zdrojov sa tak stáva čoraz dostupnejšie. Overovanie získaných vedomostí pozostáva z kontrolných úloh, kde na základe odpovedí zistíme, do akej miery študent pochopil text resp. má vedomosti, o ktorých sa v texte hovorí. Keby sme mali nástroj, ktorým by sme vedeli z textu vygenerovať faktické otázky na overenie vedomostí študenta, proces vzdelávania by sa výrazne automatizoval.

Tvorba otázok z učebného textu bola zaradená medzi úlohy patriace pod oblasť spracovania prirodzeného jazyka [6]. V porovnaní s príbuznými úlohami sa radí k náročnejším, keďže sa vyžaduje transformácia textu oboma smermi: z prirodzeného jazyka do jazyka strojov a aj naspäť. Najskôr musí byť vstupný text transformovaný do jazyka strojov - oblasť porozumenia prirodzeného jazyka (angl. Natural Language Understanding) a následne sa vytvorené otázky zobrazia v prirodzenom jazyku – oblasť generovania prirodzeného jazyka (angl. Natural Language Generation) [4]. Vstupom aj výstupom je teda text v prirodzenom jazyku.

V tomto príspevku nadväzujeme na náš predchádzajúci výskum [2] [3], kde sme opísali súčasný stav v oblasti automatického generovania otázok na základe analýzy vety. Zameriavame sa na anglický jazyk, keďže je tu možnosť porovnania sa s podobnými prácami. Náš prístup je možné prispôbiť aj na ďalšie jazyky, ale predpokladáme, že kvalita otázok bude slabšia, keďže nástroje na anotáciu textu sú najspôhlivejšie práve pre anglický jazyk. Za najviac perspektívne prístupy súčasnosti sa javia metódy založené na pravidlách a vzoroch v kombinácii s prístupmi strojového učenia. V nasledujúcej sekcii stručne zhrnieme poznatky a zistenia z aplikovania týchto prístupov, ich hlavné nedostatky a ďalej predstavíme naše riešenie na zlepšenie procesu.

2 Využívanie štruktúry vety pri generovaní otázok

Automatizované generovanie otázok sa stalo populárnou oblasťou. Prispeli tomu možnosti syntaktickej a sémantickej analýzy textu, ktoré sú prístupné pomocou viacerých nástrojov z univerzitného prostredia (napr. skupina nástrojov zo Stanfordskej univerzity¹ alebo strojovo-čitateľný slovník Wordnet²). Tie dokážu poskytnúť množstvo informácií o štruktúre textu, napríklad určiť hranice viet a slov, určiť slovné druhy, vytvoriť syntaktický strom vety či identifikovať názvoslovné entity. Získané informácie o štruktúre vety a sémantické slov poskytujú perspektívny základ pri analýze textu a keďže sa ich presnosť neustále zlepšuje, použitie je perspektívne. Veľa súčasných prístupov vychádza z metód založených na pravidlách a vzoroch (napr. [1] [4] [5]). V dizertačnej práci [4] využívajú množinu transformačných pravidiel, ktoré postupne aplikujú na vety vstupného textu na základe štruktúry vety. Najskôr pravidlami zjednodušuje zložité vety a následne zjednodušené deklaratívne vety transformujú na opytovacie vety (otázky). V [1] tiež využívajú na transformáciu štruktúru vety, ale samotná transformácia sa realizuje v jednom kroku. Pomocou orezania syntaktického stromu vety sa vytvorí kostra vety a aplikuje sa vzor, ktorý vete vyhovuje. Spoločným prvkom prístupov je využívanie štruktúry viet, ale obidva prístupy zdieľajú spoločný problém, ktorým je náročná rozšíriteľnosť vyžadujúca manuálnu tvorbu pravidiel a vzorov. Aj v [5] preukázali, že rozširovaním vzorov je možné vygenerovať kvalitnejšie otázky, ale počet vzorov výrazne narástol.

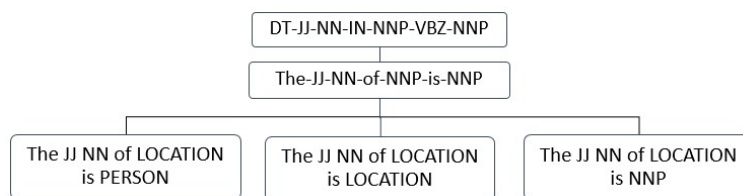
3 Generovanie otázok pomocou vetných vzorov

Vetné vzory v opisovaných prácach obsahujú súčasne slovné druhy aj názvoslovné entity. Vzhľadom na veľkú rozmanitosť viet (počet kategórií jednotlivých tokenov) je pri ich kombinovaní potrebný veľký počet vzorov, lebo každý pokryje len obmedzenú množinu viet. Preto v našom návrhu využívame kombináciu viacerých jednoduchých vzorov, ktoré uchováame v hierarchii. Jeden vetný vzor tvorí sekvenciu značiek tokenov na určitej úrovni, napríklad sekvencia slovných druhov alebo sekvencia názvos-

¹ <http://stanfordnlp.github.io/CoreNLP/>

² <http://wordnetweb.princeton.edu/perl/webwn>

lovných entít. Abstraktnejší vzor (vyššie v hierarchii) pokryje väčší počet viet, ale poskytuje všeobecnejšie informácie o vete v porovnaní s konkrétnymi vzormi, a preto aj otázky ním vytvorené sú viac všeobecné. Na najvyššej úrovni hierarchie sú abstraktné vzory reprezentujúce len syntaktickú štruktúru vety a tie sa ďalej rozlišujú na konkrétnejšie vzory, s ktorými sú prepojené (Obr. 1).



Obr. 1 Hierarchia vzorov.

Pre vetu, ktorá vyhovuje vzoru na vyššej úrovni (napr. *The capital city of Slovakia is Bratislava*) môžeme aplikovať vzor, ktorý vyhovuje nielen slovným druhom (prvý uzol hierarchie), ale má aj zhodu v ďalších tokenoch (napr. konkrétny typ predložky alebo názvoslovnej entity). Generovanie otázok je na základe tejto kombinácie, čiže keby sme uvažovali o vete, ktorá má rovnaký vzor na úrovni slovných druhov, ale rozdielne typy entít (napr. *The current president of Slovakia is Andrej Kiska*) pomocou špecifickejších vzorov vieme rozlíšiť typ otázky, ktorá sa má vytvoriť (posledný token má značku osoba, nie lokalita). Týmto sme zabezpečili efektívnejšie vyhľadanie vzoru postupne od všeobecnejších ku špecifickým a umožnili vzory rozširovať s nižšou výpočtovou zložitou ich mapovania – v prvom kroku sa vyberie podmnožina vzorov spĺňajúca základné kritériá a tieto sú následne aplikované v procese tvorby otázok. Zároveň je možné v budúcnosti vzory dopĺňať o ďalšie parametre (napr. sémantické kategórie významových slov) a tým zlepšovať pokrytie rôznych viet.

Druhým rozšírením využívania vzorov je zakomponovanie metód strojového učenia pri vyhľadávaní a odvodzovaní nových vzorov. K dispozícii máme jednak informácie o pozícii jednotlivých tokenov, slovných druhov a názvoslovných entít a zároveň databázu transformačných vzorov (pravidlá, ako sa majú deklaratívne vety zmeniť na otázky). Výber pravidiel na tvorbu otázky sa uskutočňuje na základe zhody resp. podobnosti medzi vzormi. Pri podobnosti sa zohľadňuje nielen počet zhodných značiek tokenov, ale aj možnosť ich vzájomnej zámény: niektoré značky tokenov (napr. podstatné meno a zámeno) sú navzájom ľahšie zameniteľné v porovnaní s inými dvojicami (napr. podstatné meno a sloveso). Do výpočtu podobnosti teda v súčasnosti zahrňujeme zhodu značiek, podobnosť značiek a nahraditeľnosť tokenov na jednotlivých úrovniach.

Značky tokenov využívame aj pri učení resp. tréňovaní. Iniciálna množina transformačných vzorov bola naučená na základe existujúcich dvojíc veta-otázka. Podobný princíp sa dá využiť aj pri vylepšovaní algoritmu použitím tzv. učenia s posilňovaním (angl. reinforcement learning), kedy na pravdepodobnosť aplikovanie konkrétneho

vzoru sčasti vplýva aj informácia a úspešnom použití vzoru v predchádzajúcich prípadoch. Ak sa vytvorí otázka, ktorá nebude akceptovaná, použitý vzor sa v budúcnosti aplikuje s menšou pravdepodobnosťou.

4 Záver

V príspevku sme načrtli súčasný stav automatizovanej tvorby otázok pomocou pravidiel a vzorov a dve rozšírenia, ktorými sa snažíme tento prístup vylepšiť. Prvým je uchovávanie vzorov v hierarchii vrátane vzťahov medzi nimi, vďaka čomu je možné vzory efektívnejšie vyhľadávať a spravovať tak väčšie množstvo rôznych typov vzorov. Ani to úplne nevyrieši rozmanitosť viet, pri ktorej je potrebné počet vzorov zväčšovať, ak chceme pokryť viac typov viet. Preto sa pokúšame využiť metódy strojového učenia, ktoré budú vzory odvodzovať a vytvárať na základe podobných čít tokenov, ktorými sú slovné druhy a názvoslovné entity. V budúcnosti plánujeme zohľadniť aj syntaktické vzťahy medzi slovami obsiahnuté v syntaktickom strome vety, sémantické kategórie významových slov a kategórie slov na základe konceptu prepojených dát.

Podakovanie: Tento článok vznikol vďaka podpore v rámci OP Výskum a vývoj pre projekt: Medzinárodné centrum excelentnosti pre výskum inteligentných a bezpečných informačno-komunikačných technológií a systémov, ITMS 26240120039, spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja, podpore Vedeckej a grantovej agentúry Slovenskej republiky, číslo grantu VG 1/0752/14 a VG 1/0646/15.

Literatúra

1. Ali, H. D. A. D. (2012). Automatic question generation: a syntactical approach to the sentence-to-question generation.
2. Blšták, M., & Rozinajová, V. (2016). Automatic Question Generation Based on Analysis of Sentence Structure. In International Conference on Text, Speech, and Dialogue (pp. 223-230). Springer International Publishing.
3. Blšták, M. & Rozinajová, V. (2015). Automatizovaná tvorba otázok z textu analýzou štruktúry viet. In WIKT 2015: 10th workshop on intelligent and knowledge oriented technologies. Proceedings. November 12-13, 2015 Košice, Slovakia. s. 49-52. ISBN 978-80-553-2271-1.
4. Heilman, M. (2011). Automatic Factual Question Generation from Text (Doctoral dissertation, Carnegie Mellon University).
5. Hussein, H., Elmogy, M., & Guirguis, S. (2014). Automatic English Question Generation System Based on Template Driven Scheme. International Journal of Computer Science Issues (IJCSI), (Vol. 11, pp 45-53).
6. Rus, V., Wyse, B., Piwek, P., Lintean, M., Stoyanchev, S., & Moldovan, C. (2012). A Detailed Account of The First Question Generation Shared Task Evaluation Challenge. Dialogue and Discourse, (pp. 177-204).

Annotation:

Leveraging Sentence Structure for Transformation into Question

In this paper we propose method approach to enhance and extend template-based method for automatic question generation. Actual approaches showed that template-based methods are perspective to solve this problem but they have some limitations. Main problems consist in ability to extend templates for various types of sentence structure and matching large amount of patterns to these sentences. We use multiple simple patterns stored in hierarchy which makes pattern matching easier. Although the number of patterns for covering various sentences will grow rapidly and these patterns must be created manually, we proposed to use machine learning for creation and deviation of new patterns. Learning leverages sentence structure parameters and semantic information about words (e.g.: part-of-speech tags, category of named entities and we also consider to take into account super sense tags and linked data concept).

Model používateľa pre jeho identifikáciu

Kamil Burda, Daniela Chudá

Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava, Slovenská republika

{kamil.burda, daniela.chuda}@stuba.sk

Abstrakt. V práci sa zaoberáme výskumom biometrických charakteristík správania sa pre mobilné zariadenia ako používateľsky prijateľnejšiu formu identifikácie a autentifikácie. Problémom je nedostatočná presnosť biometrických systémov pre účely identifikácie a autentifikácie používateľa, ako aj vonkajšie vplyvy (napr. polohy tela používateľa pri ovládaní zariadenia) v dôsledku mobility používania mobilných zariadení, ktoré môžu ďalej znižovať presnosť systémov. Cieľom práce je vytvoriť taký model používateľa, ktorý sa dokáže vysporiadať s vonkajšími vplyvmi, teda udržať presnosť pri rôznych vonkajších vplyvoch. Doposiaľ sa nám podarilo zistiť skutočnosť, že biometrické charakteristiky tlaku a času vykonania gesta na dotykových obrazovkách sa líšia pre jedného používateľa v rôznych telesných polohách.

Typ príspevku: Doktorandské sympóziu

Kľúčové slová: behaviorálne biometriky, modelovanie používateľa, vonkajšie vplyvy

1 Úvod

Používatelia mobilných zariadení, najmä smartfónov, často nedbajú na svoju bezpečnosť napr. pri uzamykaní smartfónu a pre zvýšené pohodlie volia slabšie heslá alebo vzory odomykania, ktoré je možné vyčítať zo stôp zanechaných na dotykovej obrazovke. Na zvýšenie bezpečnosti pri zachovaní úrovne používateľskej prívetivosti identifikácie a autentifikácie používateľa dokážeme využiť biometrické charakteristiky správania sa (behaviorálne biometriky, ďalej len „biometriky“), t.j. vzory správania sa jedinečné pre každého používateľa.

Biometrická identifikácia a autentifikácia používateľa je len jednou z možných úloh, ktoré prostredníctvom biometriky dokážeme realizovať. Vo všeobecnosti hovoríme, že systém modeluje používateľa na základe biometriky, teda vytvára si model používateľa [5]. Na realizáciu nami stanovenej úlohy v systéme pomocou biometriky (ako napr. identifikácia a autentifikácia) systém potrebuje najprv *zaznamenať* dáta – napr. poloha prstu na dotykovej obrazovke alebo tlak vyvíjaný na dotykovú obrazovku. Systém následne dáta *predspracuje* – napr. dáta rozdelíme na vzorky (napr. gesta). Z predspracovaných

dát získame biometriky (priemerný tlak na dotykovú obrazovku, čas vykonania gesta, a pod.), u ktorých predpokladáme dobrú rozlišovaciu schopnosť pre daný typ úlohy. Získané bi metriky pri prvotnom zaznamenávaní predstavujú *šablóny*. Ďalšie získané bi metriky systém použije na *porovnávanie* s uloženými šablónami a vyberie šablónu s najlepšou zhodou. Na porovnávanie sa využívajú vzdialenostné metriky, štatistické metódy alebo metódy strojového učenia (k najbližších susedov, mechanizmus podporných vektorov a pod.).

Biometrický systém nedokáže vždy vybrať správnu šablónu. Ako mieru úspešnosti systému najčastejšie používame mieru chybného prijatia (FAR), mieru chybného odmietnutia (FRR) alebo mieru, pri ktorej FAR a FRR sú rovnaké (EER), ktorú dosiahneme nastavením prahovej hodnoty pre vybranú metódu porovnávania šablón.

Vedecké štúdie v oblasti biometrie smartfónov sa zaoberali predovšetkým výskumom biometrických a ich presnosťou v systémoch za účelom autentifikácie používateľa. V Tab. 1 sa nachádza prehľad biometrických kategorizovaných podľa činností.

Tab. 1 Prehľad skúmaných biometrických

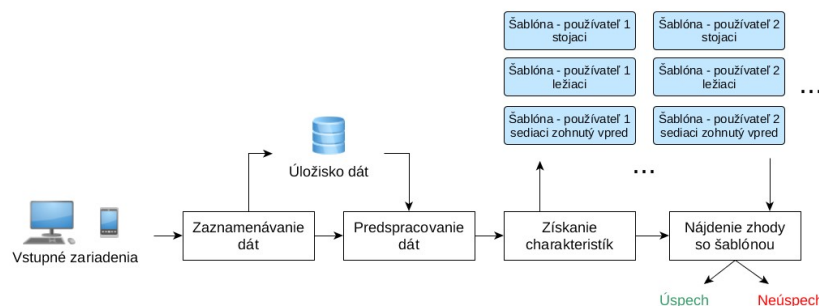
| Skupina biometrických (činnosť) | Často získavané bi metriky | EER |
|--|---|-------|
| Dynamika stláčania grafických objektov na virtuálnej klávesnici [1][7] | Čas stlačenia klávesy, medzi stlačeniami kláves, priemerný tlak na obrazovku | 3-12% |
| Gestá na dotykovej obrazovke (ťah, priblíženie, a pod.) [4] | Čas vykonania gesta, začiatkový a koncový bod gesta, priemerný tlak, smer gesta | 4-14% |
| Chôdza [6] | Štatistiky z akcelerometra (priemer, minimum, maximum), dĺžka jedného cyklu chôdze | 8-28% |
| Gestá s pohybom smartfónu (prijatie hovoru, dvíhanie smartfónu) [2][3] | Štatistiky z akcelerometra a gyroskopu, dĺžka jedného cyklu chôdze, podobnosť získaného a šablónového pohybu, | 8-20% |

Na základe chybovosti EER v štúdiách je možné skonštatovať, že pre úlohy biometrickej identifikácie a autentifikácie neposkytujú bi metriky dostatočnú presnosť. Vzhľadom na povahu používania mobilných zariadení vplyvajú na presnosť vonkajšie vplyvy ako napr. polohy tela používateľa (sediaci, stojaci, ležiaci) alebo prostredie (exteriér, interiér). Cieľom práce je navrhnúť, implementovať a overiť taký model používateľa, ktorý sa dokáže prispôbiť vonkajším vplyvom, t.j. dokáže udržať dostatočnú presnosť pri rôznych vonkajších vplyvoch.

2 Model používateľa prispôbený vonkajším vplyvom

Na Obr. 1 je znázornený všeobecný proces modelovania používateľa pre identifikáciu a autentifikáciu. Ako prvotný návrh riešenia pre vysporiadanie sa s vonkajšími vplyvmi je definovanie samostatných šablón pre jednotlivé vonkajšie vplyvy pre každého používateľa zvlášť. Problémom tohto riešenia je veľmi veľké množstvo šablón, pre ktoré

je potrebné získať potrebné vzorky. Ďalším problémom je možná podobnosť niektorých šablón navzájom.



Obr. 1 Prvotný návrh modelu používateľa pre identifikáciu a autentifikáciu so zohľadnením vonkajších vplyvov

Pre riešenie problémov s navrhnutým modelom sa potrebujeme zaoberať jednotlivými biometrikami – či sa menia alebo nemenia vzhľadom na rôzne vonkajšie vplyvy. Podobnosť jednej biometrie medzi vonkajšími vplyvmi uskutočňujeme porovnávaním párov hodnôt pre jednotlivé šablóny pre jedného používateľa. Podobnosť vyhodnocujeme pomocou štatistického testu (v našom prípade *t*-testu), pričom si stanovíme prah podobnosti a počet vonkajších vplyvov, kedy je daná biometrika podobná.

Na základe riešenia problému sme vykonali prvý experiment so 43 účastníkmi, v ktorom sme skúmali, aké charakteristiky sa menia pri rôznych polohách tela pri vykonávaní jednoduchých ťahov na dotykovej obrazovke smartfónu. Získali sme dovedna 11 biometrických údajov týkajúcich sa tlaku na dotykovú obrazovku, trajektórie gesta, koncových bodov gesta a času vykonania gesta. Na základe experimentálnych výsledkov sme zistili, že priemerný a maximálny tlak na dotykovú obrazovku rozlišovalo polohy pre viac ako 50% používateľov, čas vykonania gesta pre viac ako 22% používateľov a časový okamih s najväčšou odchýlkou trajektórie gesta od vzdialenosti koncových bodov gesta pre viac ako 18% používateľov.

3 Záver

V práci sa zaoberáme výskumom biometrických charakteristík správania sa pre mobilné zariadenia za účelom identifikácie a autentifikácie používateľa. Nedostatočnú presnosť biometrických systémov dokážeme znížiť získaním ďalších biometrických údajov z doposiaľ málo preskúmaných činností. Väčší problém pri používaní mobilných zariadení však predstavujú vonkajšie vplyvy (napr. polohy tela), ktoré takisto môžu znižovať presnosť systémov.

Na základe problému s vonkajšími vplyvmi sme navrhli model používateľa, ktorý sa dokáže vysporiadať s vonkajšími vplyvmi. Doposiaľ sa nám podarilo zistiť skutočnosť, že biometricky tlaku a času vykonania gesta na dotykových gestách sa líšia pre jedného používateľa v rôznych telesných polohách.

Podakovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektu Prispôsobovanie prístupu k informačným a vedomostným artefaktom založené na interakciách a kolaborácii v prostredí webu, grant No. VG 1/0646/15, Informačné správanie sa človeka v digitálnom priestore, grant APVV-15-0508.

Literatúra

1. Chang, T.-Y. et al.: A graphical-based password keystroke dynamic authentication system for touch screen handheld mobile devices. *Journal of Systems and Software*. May 2012. Vol. 85, no. 5, p. 1157–1165. DOI 10.1016/j.jss.2011.12.044.
2. Conti, M. et al.: Mind How You Answer Me!: Transparently Authenticating the User of a Smartphone when Answering or Placing a Call. In: *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*. New York, NY, USA: ACM, 2011. p. 249–259. ASIACCS '11. ISBN 978-1-4503-0564-8.
3. Feng, T. et al.: Investigating Mobile Device Picking-up motion as a novel biometric modality. In: *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. September 2013. p. 1–6.
4. Frank, M. et al.: Touchalytics: On the Applicability of Touchscreen Input as a Behavioral Biometric for Continuous Authentication. *IEEE Transactions on Information Forensics and Security*. January 2013. Vol. 8, no. 1, p. 136–148. DOI 10.1109/TIFS.2012.2225048.
5. Modi, S.K.: *Biometrics in Identity Management: Concepts to Applications*. 1 edition. Boston: Artech House, 2011. ISBN 978-1-60807-017-6.
6. Nickel, C. et al.: Authentication of Smartphone Users Based on the Way They Walk Using k-NN Algorithm. In: *2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*. July 2012. p. 16–20.
7. Saevanee, H., Bhattarakosol, P.: Authenticating User Using Keystroke Dynamics and Finger Pressure. In: *6th IEEE Consumer Communications and Networking Conference, 2009. CCNC 2009*. January 2009. p. 1–2.

Annotation:

User Model for Identification

Our work deals with the research of behavioral biometrics for mobile devices as a more user-convenient form of user identification and authentication. One of the problems in the research area is the relatively low accuracy of biometric systems for the purposes of identification and authentication, as well as external factors (such as body postures of a user) due to the mobility of the users. The goal of our work is to create a user model that can cope with the external factors and maintain a reasonable level of accuracy given the external factors. So far we have discovered that touch pressure and length of simple swipes on a smartphone touch screen vary in different body postures in each user individually.

Automation and visualization in ontological engineering leveraging on background models

Marek Dudáš, Vojtěch Svátek

Fakulta informatiky a statistiky
Vysoká škola ekonomická
Praha 4, Česká republika

{marek.dudas,svatek}@vse.cz

Abstract. The common way of OWL ontology development for semantic web is to create and work with the ontologies directly in the RDFS/OWL language. That might make the task harder than it could be, since OWL allows to encode the same real world situation using different combinations of language constructs. We propose splitting the ontology development into two steps. First, the relevant entities and relationships are described in an ontological background model and then the encoding style is chosen for each entity in the second step. Finally, the seed of an OWL ontology is generated automatically from the background model. The PhD thesis focuses on development of visualization and transformation methods and their implementations as graphical tools that will allow to test the proposal with users.

Contribution type: PhD Symposium

Keywords: ontology engineering, OWL, ontological background models

1 Introduction

Ontologies used as data schemas to describe the data on the semantic web are its essential component. The common way of ontology development is to work with the ontologies directly in the Web Ontology Language¹ (OWL). That might make the task harder than it could be, since OWL allows to encode the same real world situation using different combinations of language constructs, following different encoding styles, which might affect the suitability of the resulting ontology for various use cases. The engineer has to deal with two problems at the same time: defining what concepts are in the modeled domain and choosing the OWL encoding style for them. We propose splitting the ontology development into two steps. First, the relevant concepts are described in an *ontological background model*, analogical to an entity-relationship diagram in database design, and then the encoding style is chosen for each entity in the second

¹ <https://www.w3.org/TR/owl2-primer/>

step. Finally the seed of an OWL ontology is generated automatically from the background model. Similar issue is at the side of the ontology user. When users want to learn how to use terms from the ontology to describe their data, their only options are to look directly at the ontology in OWL or ontology documentation, which is often not provided and rarely visual. That might make the task difficult, since the source OWL representation of an ontology shows what terms *can* be used, but not *how*, in what combinations, they should be used. The “how to use” visualization could be achieved by summarizing a dataset where the ontology is already in use, leading to basically learning by example. The PhD thesis focuses on development of visualization and transformation methods and their implementation that will allow testing the proposal with users. The idea is illustrated by Figure 1. Simply said, the goal is making the work of ontology engineers and users easier by adding another interface layer between them and the ontology in its source form.

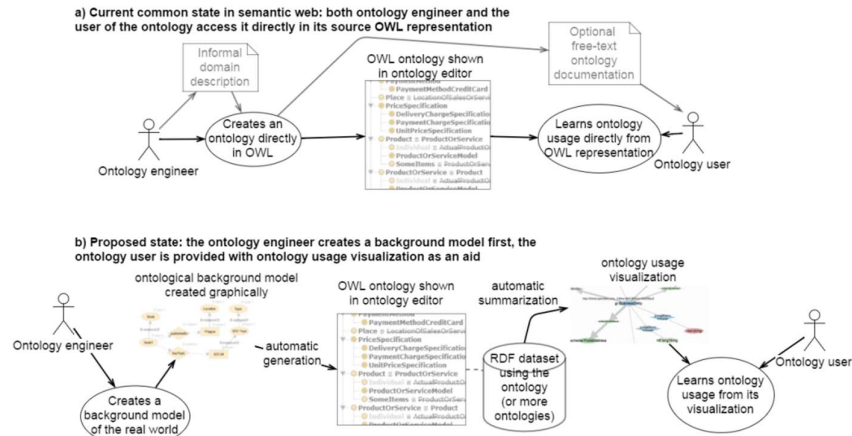


Fig. 6. Thesis proposal (b) compared to current state in ontology engineering (a)

2 State of the art

The problem of heterogeneity of ontologies is targeted by a whole research area of ontology mapping [6]. It aims at enabling usage of a combination of different ontologies, however, it is not concerned with the encoding style heterogeneity. Meta-modeling approaches might allow abstracting from the OWL encoding differences. PURO ontological background models (OBM) [7] allows modeling a part of reality in a representation that relaxes some of the constraints imposed to OWL by its description logic grounding and can be mapped to different OWL encoding styles. A similar meta-modeling approach offers OntoClean [3], which however only focuses on classes in a taxonomy and its intended usage is for coherence testing. OntoUML [1] is a version of UML for conceptual modeling where the modeling primitives are grounded in concepts of a foundational ontology. That allows validation of the models against syntactical

errors and application of ontological design patterns. OLED [2], a graphical editor for OntoUML, allows to transform it into OWL fragments. The transformation is hard-coded and each OntoUML element has its single OWL counterpart – encoding style heterogeneity is not considered. PURO language seems to be the most promising way of representing of the ontological background models thanks to it being mentally close to OWL and was therefore chosen for the proposal implementation.

Several dataset summarization tools usable to study ontology usage in a dataset exist. The main problem with them is that all of them stayed in a very experimental stage of development and are not publicly available. The same principles as are proposed in the thesis uses maps of ontology usage [5]. ExpLOD [4] offers a more complex approach based on bisimulation contraction. The result is a node-link visualization similar to what we propose but more accurate: showing a combination of links that reportedly exist in the dataset while we show combinations of links that possibly exist. Our visualization might be on the other hand more intuitive as it shows types of instances directly as node labels while ExpLOD shows types as separate nodes which might lead to clutter.

3 Achievements so far

Visual authoring of ontological background models in a web-based tool PURO Modeler² and their transformation to OWL in OBOWLMorph³ has been developed and preliminary evaluated with users. A tool for visualization of ontology usage as combinations of types and properties in a graph, LODSight,⁴ has been developed, but has not been tested with users yet. An important starting point for the research lies in ontology visualization. Therefore, a comprehensive survey of ontology visualization tools, including an updated classification of visualization methods, has been done.

4 Evaluation

We have evaluated PURO Modeler and OBOWLMorph concept with a group of ten students with basic knowledge about OWL from a course of ontology engineering. The aim was to compare ontology engineering in our tools to common ontology editor Protégé.⁵ The hypothesis was that PURO-started development allows beginner-users to create an ontology appropriately covering the domain with less effort than common ontology development in Protégé. The students were assigned to create a PURO model and an ontology in Protégé according to a textual description of the model and were given a questionnaire afterwards. Very brief overview of the evaluation results follows: modeling in PURO is a little bit slower and more error prone (partially due to less strict UI), however leads to better coverage of the domain. Both the time consumption and

² <http://protegeserver.cz/puromodeler/>

³ <http://protegeserver.cz/puromodeler/OBOWLMorph/>

⁴ <http://lod2-dev.vse.cz/lodsight-v2/>

⁵ <http://protege.stanford.edu>

number of errors is quite similar in OWL and PURO. According to the questionnaire, students prefer PURO Modeler over Protégé. They consider PURO to be rather easy to learn and did not hesitate much about which PURO construct to use for each entity. Given that PURO Modeler and OBOWLMorph are at early stage of development and their UI is not very user friendly yet, the results are quite encouraging. LODSight has been so far tested only from the technical point of view. Evaluation with users is planned as future work.

5 Conclusion

We have proposed a method for starting ontology development from ontological background models exploiting the existing PURO language. The method has been implemented and evaluated with users. The evaluation suggests the tools are usable, but need much improvement. A tool for visualization of existing ontology usage, useful for both ontology developers considering reuse of existing ontologies and ontology users, has been developed but not yet tested with users. The tool builds on existing methods, adds new features and improves ease of use compared to similar existing tools. The future research will aim at evaluation of the whole framework with users, improvements based on it, and integration of a semi-automated way of reusing concepts from existing ontologies into the PURO-started ontology development process.

Acknowledgment: This research is supported by UEP IGA F4/28/2016.

References

1. Albuquerque, A., & Guizzardi, G. (2013, May). An ontological foundation for conceptual modeling datatypes based on semantic reference spaces. In IEEE 7th International Conference on Research Challenges in Information Science (RCIS) (pp. 1-12). IEEE.
2. Barcelos, P. P. F., dos Santos, V. A., Silva, F. B., Monteiro, M. E., & Garcia, A. S. (2013, October). An Automated Transformation from OntoUML to OWL and SWRL. In Ontobras (pp. 130-141).
3. Guarino, N., & Welty, C. A. (2009). An overview of OntoClean. In Handbook on ontologies (pp. 201-220). Springer Berlin Heidelberg.
4. Khatchadourian, S., & Consens, M. P. (2010, May). ExpLOD: summary-based exploration of interlinking and RDF usage in the linked open data cloud. In Extended Semantic Web Conference (pp. 272-287). Springer Berlin Heidelberg.
5. Kinsella, S., Bojars, U., Harth, A., Breslin, J. G., & Decker, S. (2008, July). An interactive map of semantic web ontology usage. In 2008 12th International Conference Information Visualisation (pp. 179-184). IEEE.
6. Shvaiko, P., & Euzenat, J. (2013). Ontology matching: state of the art and future challenges. IEEE Transactions on knowledge and data engineering, 25(1), 158-176.
7. Svátek, V., Homola, M., Kluka, J., & Vacura, M. (2013, June). Mapping structural design patterns in OWL to ontological background models. In Proceedings of the seventh international conference on Knowledge capture (pp. 117-120). ACM.

Vplyv individuálnych vlastností používateľov na výsledky používateľských štúdií

Patrik Hlaváč, Mária Bieliková

Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava, Slovenská republika

{patrik.hlavac, maria.bielikova}@stuba.sk

Abstrakt. Používateľské štúdie v oblasti Webu sú založené na niektorých metri-
kách, ale otázkou je, ako tieto metriky aplikovať na väčšiu skupinu používateľov
(účastníkov štúdie). Keď uvažujeme, že každý účastník má rôzne vlastnosti, skú-
senosti a zručnosti, očakávame tiež, že výsledky štúdií v rovnakom prostredí
budú mať rôzne hodnoty. Zameriavame výskum na ukázanie, že z kvantitatív-
nych štúdií môžeme získať presnejšie výsledky, ak zoberieme do úvahy aj jed-
notlivé vlastnosti o osobných charakteristikách účastníkov.

Typ príspevku: Doktorandské sympóziu

Kľúčové slová: kvantitatívne štúdie, individuálne rozdiely, používateľský záži-
tok

1 Úvod a motivácia

Používateľské štúdie pomáhajú zamerať sa na konkrétny problém, napríklad pri návrhu dizajnu alebo overovaní použiteľnosti. Niekedy je vhodné využiť väčšiu vzorku účastníkov. Keď uvažujeme, že každý používateľ má rôzne schopnosti, zručnosti a skúsenosti, očakávame tiež, že výsledky testovania sa budú líšiť. Výsledky môžu byť ovplyvnené viacerými vplyvmi, niektoré z nich už boli identifikované.

Používateľské štúdie delíme na kvalitatívne a kvantitatívne. Zatiaľ čo kvalitatívne štúdie zvyčajne pozostávajú z interakcie účastníka v danom prostredí za účasti mode-
rátor, ako dôležitého sprostredkovateľa, kvantitatívne štúdie sú zvyčajne vykonávané bez neho a teda bez hlbšej analýzy konkrétneho používateľa. Pri kvantitatívnych štúdiách sa presúvame zo špecifických detailov ku generalizovanej informácii pre celú skupinu účastníkov.

Vyhodnotenie používateľského testovania môže byť presnejšie s dodatočnou infor-
máciou o účastníkových zručnostiach, ako napríklad Webová alebo Počítačová gramot-
nosť. Pojem Webovej gramotnosti sa počas rokov menil a v súčasnosti ho vystihujú
aspekty: čítanie, písanie a zúčastnenie sa, niekedy označované ako oblasti: skúmanie,

tvorenie a spájanie¹. Spolu potom obsahujú V našej práci sa snažíme odhaliť základné vzťahy medzi Webovou gramotnosťou a prácou vo webovom prostredí. Práve rozdiely v používaní Webu jednou zo skupín (účastníci s vysokou alebo nízkou Webovou gramotnosťou) nás môžu nasmerovať k lepšiemu porozumeniu základných princípov.

2 Súvisiace práce

Naša pozornosť sa upriamuje na štúdie, ktoré vyhodnocujú interakciu používateľa so zameraním na individuálne rozdiely účastníkov v kvantitatívnych štúdiách. Napriek dlhodobému výskumu v oblasti použiteľnosti, sa tejto téme venuje len okrajovo.

Individualitu, ako unikátnosť voči ostatným, popisujú mnohé psychologické a medicínske štúdie [5]. Ukazuje sa, že práve individualita má veľký vplyv na výsledky štúdií. Základné vplyvy tvoria psychologické črty (neurotizmus, extravergia, otvorenosť, prívetivosť, svedomitosť) [7] zväčša modelované pomocou dotazníkov. Počas mnohých rokov výskumu vznikli nástroje, schopné identifikovať psychologické črty jednoduchšie, napríklad z formy písaného textu. Postupne sa rozvíja výskum vplyvu veku a pohlavia, niekedy obohatené o skúmanie skúsenosti v danej oblasti či vzdelania [6]. Testovanie vplyvu pohlavia neukazuje jednoznačný vplyv. Niekedy sa nevyskytujú odlišnosti vnímania kvality služieb či informačnej kvality [7], ale ukazujú sa rozdiely vnímania kvality obsahu [3]. Výsledky sú podľa všetkého závislé od ďalších faktorov. Pri testovaní žiakov sa rozdiely medzi pohlaviami ukázali napríklad pri vnímaní grafickej a textovej informácie, ktoré sa pripisujú lepším jazykovým vlastnostiam dievčat. Objavujú sa tiež rozdiely vo vyhľadávacích vzoroch [1]. Ďalšia štúdia [2] využila sebahodnotenie používateľov v otázke Internetovej gramotnosti v závislosti od doby používania sociálnych médií. Z iných štúdií vieme, že využívanie subjektívneho hodnotenia nie je veľmi presné a zaoberáme sa nesubjektívnym vyhodnotením. Testovanie Webovej alebo Digitálnej gramotnosti dnes poskytujú najmä spoločnosti, ktoré vedú aj vzdelávanie v týchto oblastiach. Forma testovania je rôzna: od aplikácií, ktoré bežia na desktopoch až po online dotazníky.

3 Otvorené problémy a výskumné ciele

Aktuálnym problémom je, že používateľské štúdie neberú do úvahy individuálne vlastnosti používateľov a teda ani Webovú gramotnosť. Naším záujmom je výskum vplyvu individuálnych vlastností účastníkov štúdií na tieto štúdie. Momentálne sa zaujíname o vzťah Webovej gramotnosti a miery do ktorej môže ovplyvniť výsledky experimentov. Cieľom je navrhnúť metódu na detekciu individuálnych rozdielov.

¹ <https://wiki.mozilla.org/Learning/WebLiteraciesWhitePaper>

4 Metóda na určovanie Webovej gramotnosti

Navrhujeme vlastné špecifické riešenie pozostávajúce zo série testov na detekciu charakteristík, konkrétne Webovej gramotnosti ako súčasti Digitálnej gramotnosti. Návrh sa zakladá na schopnosti správne určiť rozhodovacie testy pre ďalšiu automatizáciu. Základný motív je porovnávanie účastníkov s vyššou gramotnosťou a účastníkov s nižšou gramotnosťou. Predpokladom je získanie informácií z interakcie účastníka na Webe. Webová gramotnosť používateľa je určovaná pomocou testu pozostávajúceho z troch častí.

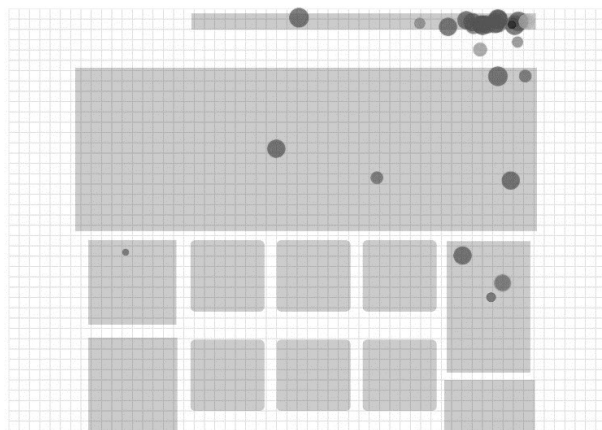
Prvá časť skúma gramotnosť explicitne pomocou dotazníka. Účastníci nehodnotia sami seba, ale odpovedajú na kvíz zložený z relevantných otázok. Túto časť sme realizovali dotazníkmi Google Forms, kde sme použili 14 otázok a 4 dostupné odpovede.

Ďalšia časť zisťuje znalosť webových ikon bežne používaných na webstránkach. Účastník dokazuje, ako dobre sa v nich vyzná, pomocou voľby vhodných ikon (napr. „menu“ alebo „poslať e-mail“) na základe otázky. Takto sme testovali 15 charakteristických webových ikon.

V tretej časti hľadáme základné vzory v hľadaní oblastí na webstránke, znova v závislosti od gramotnosti účastníkov. Účastníci majú za úlohu označiť miesto na obrazovke, kde očakávajú výskyt dopytovaného elementu. Plánujeme porovnať závislosti označenej pozície elementu a určenej gramotnosti. Miesto reálnych stránok im zobrazujeme vizuálne upravenú schému, ako na obrázku 1. Tieto schémy obsahujú informáciu o type stránky „značka elektroniky“, „fakultná webstránka“, „filmová databáza“ a pokyn, aký prvok majú hľadať (napr. „nákupný košík“ alebo „vyhľadávanie“).

Pilotný experiment sme vykonali v UX Labe na FIIT STU za pomoci okulografu Tobii TX300. Infraštruktúra [4] umožňuje zber informácií z webkamery, obrazovky, okulografu a použitého webového prehliadača.

Testovanie používateľov prebiehalo na diaľku, v stanovenom poradí testov a náhodnom poradí úloh v nich.



Obr. 1 Ukážka vzoru s mriežkou a pozíciami označenými účastníkmi.

5 Záver

Webovú gramotnosť považujeme za kľúčový aspekt dnešnej interakcie človeka s počítačom. Zatiaľ existuje iba obmedzený počet prístupov na určovanie Webovej gramotnosti, preto navrhujeme vlastnú metódu. Aktuálne nastavenie experimentov ukázalo, že je treba zaviesť vhodnejšie a detailnejšie sledovanie interakcie účastníka pre získanie relevantných výsledkov. V ďalšej práci sa teda zameriame na automatizované vyhodnocovanie stupňa Webovej gramotnosti na základe detailnejšej informácie o interakcii účastníkov v danom prostredí.

Podakovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektov VEGA 1/0646/15 a HIBER APVV-15-0508.

Literatúra

1. Hsieh, T. Y., & Wu, K. C. (2015). The Influence of Gender Difference on the Information-Seeking Behaviors for the Graphical Interface of Children's Digital Library. *Universal Journal of Educational Research*, 3(3), 200-206.
2. Len-Ríos, M. E., Hughes, H. E., McKee, L. G., & Young, H. N. (2015). Early adolescents as publics: A national survey of teens with social media accounts, their media use preferences, parental mediation, and perceived Internet literacy. *Public Relations Review*.
3. Lu, H.-P. & Lee, M.-R., 2010. Demographic differences and the antecedents of blog stickiness. *Online Information Review*, 34(1), pp. 21-38.
4. Móro, R., Daráz, J., & Bielíková, M. (2014). Visualization of Gaze Tracking Data for UX Testing on the Web. In HT (Doctoral Consortium/Late-breaking Results/Workshops).
5. Pietilä, S., Björklund, A., & Bülow, P. (2013). 'We are not as alike, as you think'sense of individuality within the co-twin relationship along the life course. *Journal of aging studies*, 27(4), 339-346.
6. Sonderegger, A., Schmutz, S., & Sauer, J. (2016). The influence of age in usability testing. *Applied Ergonomics*, 52, 291-300.
7. Zha, X., Zhang, J., Yan, Y., & Xiao, Z. (2014). User perceptions of e-quality of and affinity with virtual communities: The effect of individual differences. *Computers in Human Behavior*, 38, 185-195.

Annotation:

Impact of Characteristics of Individuals on Evaluating the Quantitative Studies

Usability studies in the web domain are based on various metrics, but the question is how to apply these metrics to evaluate a larger group of people. When we consider that every user has different qualities, skills and experiences, we could expect that the results of testing of same scenarios will be different. We aim our research to show that quantitative studies can provide more accurate results if we work with information about personal characteristics of participants. We have already conducted a preliminary controlled experiment on a small sample of participants, which explores influence of a Web literacy.

Metóda kombinácie predikčných modelov na spresnenie predikcie spotreby elektrickej energie

Marek Lóderer, Viera Rozinajová

Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava, Slovenská republika

{marek_loderer, viera.rozinajova}@stuba.sk

Abstrakt. Rozvojom a zavádzaním inteligentných meračov (*Smart Meters*) dochádza k zhromažďovaniu nových typov dát. Tieto dáta poskytujú informácie o aktuálnej spotrebe a priebehoch odberov jednotlivých odberno-odovzdávacích miest. Získané dáta otvárajú možnosti na vytváranie nových modelov s cieľom spresniť predikciu spotreby elektrickej energie. Vzhľadom na obmedzené možnosti výroby a uskladňovania vyrobenej el. energie, je tento problém stále vysoko aktuálny. V práci je predstavená metóda kombinujúca výsledky predikcie viacerých predikčných modelov. Uvedený prístup sa v literatúre nazýva *Ensemble Learning*. Dôležitou súčasťou metódy je spôsob kombinácie čiastkových výsledkov do finálnej predikcie. Tento zložitý numerický problém riešime pomocou biologicky inšpirovaných algoritmov, ktoré dokážu v konečnom čase a pri pomerne nízkych výpočtových nárokoch, poskytnúť optimálne riešenie.

Typ príspevku: Doktorandské sympóziu

Kľúčové slová: časové rady, predikcia, učenie súborom metód, biologicky inšpirované algoritmy

6 Úvod

V súčasnosti existuje niekoľko desiatok rôznych predikčných metód, určených na prácu s časovými radmi. Vo všeobecnosti môžeme tieto metódy rozdeliť do troch hlavných skupín [7]:

- *tradičné metódy* (regresia, viacnásobná regresia, exponenciálne vyrovňovanie)
- *modifikované tradičné metódy* (adaptívne metódy, stochastické metódy, autoregresný model ARMA a ARIMA, regresia založená na podporných vektoroch)
- *metódy umelej inteligencie* (evolučné algoritmy, fuzzy logika, neurónové siete, znalostné expertné systémy a iné.)

Každá z uvedených metód má svoje výhody a nevýhody. Vzhľadom na komplexný problém predikcie časových radov, resp. predikcie spotreby el. energie, je zložité zvoliť

jednu konkrétnu metódu, ktorá dokáže vždy poskytnúť správny výsledok. Riešením je prístup založený na kombinácii viacerých predikčných modelov.

7 Učenie súborom metód

Učenie súborom metód (*Ensemble Learning*) je jedným z prístupov z oblasti strojového učenia, ktoré môže byť definované, ako proces pozostávajúci z trénovania a kombinácie rozličných modelov, ktorých úlohou je vyriešiť zadaný problém [6]. Podobný prístup môže byť pozorovaný v ľudskom správaní. Príkladom môže byť parlament alebo senát, kde sa pri prijímaní dôležitého rozhodnutia, berú do úvahy názory viacerých odborníkov.

Učenie súborom metód môže byť použité na zlepšenie výsledkov zhukovacích, klasifikačných i predikčných modelov [8, 9]. Uvedený proces je silne závislý na troch hlavných komponentoch. Prvý komponent zabezpečuje generovanie rozličných predikčných modelov. Druhý komponent rozhoduje o tom, ktoré vygenerované modely sa ponechajú a ktoré budú kvôli nepostačujúcim výsledkom odstránené zo základného súboru. Posledný komponent je zodpovedný za integráciu jednotlivých modelov s cieľom spresniť konečný výsledok predikcie.

7.1 Generovanie sady predikčných modelov

Pod generovaním sady predikčných modelov sa rozumie natrénovanie jednotlivých modelov na množine trénovacích dát. Pri trénovaní modelov sa využíva heterogénny prístup (jednotlivé modely sú trénované na rovnakých datasetoch). Použité predikčné modely (Tab. 1) sa líšia spôsobom výpočtu a nárokmi na veľkosť trénovacieho okna.

Niektoré predikčné modely môžu byť rozšírené o externé faktory, ako napríklad predpovede počasia, ktoré dokážu spresniť výslednú predikciu.

Tab. 1 Zoznam použitých predikčných modelov v súbore. Modely môžu byť rozdelené do troch skupín: TM – tradičné metódy (regresné modely), MTM – modifikované trad. metódy (modely založené na analýze časových radov) a AI – modely založené na umelej inteligencii.

| | Názov modelu | Typ modelu | Zahrnutie externých faktorov |
|---|---|------------|------------------------------|
| 1 | Viacnásobná lineárna regresia | TM | áno |
| 2 | Dopredná neurónová sieť | AI | áno |
| 3 | Rekurentná neurónová sieť | AI | áno |
| 4 | Hlboká neurónová sieť | AI | áno |
| 5 | Regresia založená na podporných vektorech | AI | áno |
| 6 | Náhodné lesy | AI | áno |
| 7 | Plávajúci priemer | MTM | nie |
| 8 | ARIMA model | MTM | nie |
| 9 | Dekompozícia časového radu a predikcia jednotlivých zložiek | MTM | nie |

7.2 Orezávanie sady modelov

Orezávanie (redukcia) sady modelov slúži na spresnenie výslednej predikcie a zníženie výpočtovej a pamäťovej náročnosti. V tomto kroku sú zo sady modelov vyučené modely s najhoršími výsledkami. Najčastejšie sa využívajú dva prístupy: rozdeľovací a vyhľadávací [6].

7.3 Integrácia

Posledným krokom v procese učenia súborom metód je spojenie výsledkov jednotlivých modelov v sade do finálneho výsledku. Na rozdiel od klasifikačných metód, kde je finálny výsledok určený na základe najčastejšej odpovede, pri regresných problémoch je integrácia výsledkov predikčných modelov komplikovanejšia. Jedna z metód, ktorá sa na tento účel využíva je vážený priemer (rovnica 1):

$$F_{final} = \frac{\sum_{i=1}^m w_i * F_i}{\sum_{i=1}^m w_i} \quad (1)$$

kde F_{final} je výsledná predikcia, F_i je predikcia i -teho modelu, m je počet predikčných modelov, ktoré vstupujú do fázy integrácia a w_i je váha i -teho modelu vo váženom priemere.

Vážený priemer v porovnaní s obyčajným priemerom, umožňuje pomocou váh zvýhodňovať a penalizovať modely na základe ich výsledkov. Nové hodnoty váh sú vypočítané na základe chýb predikcie jednotlivých modelov. Chyba predikcie je vypočítaná ako priemerná absolútna percentuálna chyba MAPE - *Mean Absolute Percentage Error* [2].

Výpočet váh môžeme charakterizovať ako optimalizačný problém s ohraničením. Na základe nášho predchádzajúceho výskumu [1] sme sa rozhodli použiť optimalizačné algoritmy: Umelá kolónia včiel [3] a Optimalizácia rojom častíc [4], ktoré dosiahli dobré výsledky a mali nízku časovú náročnosť v porovnaní s ostatnými testovanými algoritmi.

8 Záver

Naším cieľom je vytvoriť metódu založenú na učení súborom metód, ktorá bude vhodná pre prácu časovými radmi a bude produkovať presné predikcie. Vytvorenú metódu ako aj naše predpoklady testujeme na dátach o spotrebe elektrickej energie. Naším zámerom je vytvárať presnejšie krátkodobé predikcie spotreby elektrickej energie pomocou učenia súborom metód a biologicky inšpirovaných algoritmov. Experimenty vykonané na reálnych dátach z inteligentných meračov potvrdzujú úspešnosť navrhnutej metódy [5]. V rámci experimentov [1] sme skúmali aj schopnosť rôznych biologicky inšpirovaných algoritmov zvýšiť predikčnú schopnosť súboru metód v situáciách, keď

sa v dátach objavujú nepredvídateľne zmeny (náhle ako aj postupné). Najlepšie výsledky dosahovali rojovo inteligentné algoritmy.

Podakovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektu VEGA 1/0752/14.

Literatúra

1. Bou Ezzeddine, A., Lóderer, M., Laurinec, P., Vrablecová, P., Rozinajová, V., Lucká, M., Lacko, P., Grmanová, G.: Using Biologically Inspired Computing to Effectively Improve Prediction Models. In *The International Journal of Hybrid Intelligent Systems*, (2016)
2. Hyndman, R.J., Koehler, A.B.: Another look at measures of forecast accuracy, In *International Journal of Forecasting*, vol. 22, no. 4, (2006), pp. 679-688
3. Karaboga, D., Basturk, B.: A powerful and efficient algorithm for numerical function optimization: artificial bee colony (abc) algorithm. In *Journal of global optimization*, (2007)
4. Kennedy, J., Eberhart, R.: Particle swarm optimization. In *Proceedings ICNN'95 - International Conference on Neural Networks*. vol. 4, (1995), pp. 1942-1948.
5. Kosková, G., Rozinajová, V., Bou Ezzeddine, A., Lucká, M., Lacko, P., Lóderer, M., Vrablecová, P., Laurinec, P.: Application of Biologically Inspired Methods to Improve Adaptive Ensemble Learning. In *NaBIC 2015. Advances in nature and biologically inspired computing*. Pietermaritzburg, South Africa, 2015, Springer, (2016), pp. 235-246.
6. Mendes-Moreira, J., Soares, C., Jorge, A.M., Freire de Sousa, J.: Ensemble approaches for regression: A survey. *ACM Computing Surveys (CSUR)*, vol. 45, no. 1, (2012), pp. 10.
7. Singh, A.K., Ibraheem, S.K., Muazzam, M.: An Overview of Electricity Demand Forecasting Techniques. In *National Conference on Emerging Trends in Electrical, Instrumentation and Communication Engineering IISTE 2013*, vol. 3, no. 3, (2013)
8. Strehl, A., Ghosh, J.: Cluster ensembles: A knowledge reuse framework for combining multiple partitions. In *Journal of Machine Learning Research*. vol. 3, (2003), pp. 583-617
9. Xiao, L., Wang, J., Hou, R., Wu J.: A combined model based on data pre-analysis and weight coefficients optimization for electrical load forecasting. In *Energy*, vol. 82, (2015)

Annotation:

Method of prediction models combination used to precise prediction of power load consumption

Ensemble learning is one of the machine learning approaches that can be defined as the process of training and combining diverse models to solve a particular computational problem [6]. Ensemble learning can be used for improving the performance of clustering, classification or prediction [8, 9]. The whole process depends on three main parts: first the way how the set of diverse models is created, second which models are eliminated from the set depending on their performance and third the way of integrating the models into final prediction. In our research we investigate different ways of constructing accurate ensemble. We focus on different weighting schemes of predictive base models, especially from the field of biologically inspired algorithms, like Artificial Bee Colony [3] or Particle Swarm Optimization [4]. Our goal is to create ensemble learning method suitable for precise short-term load predictions.

Analýza dát za účelom zlepšenia konkrétného firemného procesu logistickej firmy

Miroslava Muchová, Ján Paralič

Fakulta elektrotechniky a informatiky
Technická univerzita v Košiciach
Letná 9/B, 042 00 Košice, Slovenská republika

{miroslava.muchova, jan.paralic}@tuke.sk

Abstrakt. Dolovanie v dátach je veľmi dôležité pre moderné riadenie logistiky, čo pomáha zlepšiť správne rozhodnutia, zvýšiť predaj, znížiť náklady a pod. V kontexte týchto rozhodnutí hrajú kľúčovú rolu nielen správne informácie a znalosti, ale aj spôsob ako ich efektívne použiť. Predkladaný článok sa zaoberá problematikou analýzy dát pre zlepšenie rozhodovania vo vybranom logistickom procese – výber vodičov na plánované dodacie trasy. Článok predstavuje naše úvodné kroky pri riešení jednotlivých podúloh pomocou rozhodovacích stromov a jednoduchých štatistických metód, ktoré sme aplikovali na dáta z konkrétnej firmy a dosiahnuté výsledky prezentujeme v tomto článku.

Typ príspevku: Doktorandské sympóziu

Kľúčové slová: analýza dát, logistika, RapidMiner

1 Úvod

Analýza dát a logistika do seba dokonale zapadajú. Logistické spoločnosti riadia často veľký tok tovaru, pričom vytvárajú množstvo dát. Tieto dáta v sebe skrývajú potenciál pre nové obchodné modely. Riadenie zásob, sledovanie zásielok a dokonca umiestnenie senzorov vo vozidlách, všetky tieto činnosti poskytujú veľké množstvo dát [1] [3]. Pre logistické podniky je ťažké uskutočňovať včasné a presné rozhodnutia na riadenie procesu a prevádzkovú činnosť logistiky [2]. Technológie dolovania v dátach a štatistické analýzy môžu pomôcť pochopiť správanie zákazníkov a vykonať zodpovedajúcu stratégiu, čím firma dokáže znížiť riziko plynúce z chybného rozhodnutia [8].

1.1 Súčasný stav problematiky

V poslednej dobe rôzne výskumné štúdie poukázali na výhody použitia veľkých dátových metód v oblasti logistiky a riadenia dodávateľského reťazca. V nami sledovanej oblasti logistiky boli publikované zaujímavé výsledky v rámci FP7 projektu Companion. Technická správa [5] popisuje okrem iného výsledky výskumu zameraného na

tvorbuprediktívnych modelovspotreby paliva pre nákladné vozidlá. Autori pre predikciu spotreby paliva využívajú metódy strojového učenia a to prostredníctvom rôznych faktorov ako napr. charakteristiky trasy, vozidla, hmotnosť nákladu, správanie vodiča pri riadení, ale aj počasie. Vytvorený model by tak mal dokázať stanoviť očakávané náklady na rôznych cestách pri plánovaní a optimalizácii trasy. Dáta pochádzali z niekoľkých rôznych zdrojov, vrátane databázy pre správu vozového parku, konfigurácie vozidla databázy, cestnej databázy a historických údajov o počasí. Autori vytvorili prediktívne modely pomocou lineárnej regresie, náhodných stromov, SVM a neurónových sietí. Najlepšiu presnosť dosahovali náhodné stromy (pri predikcii v minútových intervaloch priemerne 21,6% chyba, pri 10-minútových intervaloch 13,1%). Výsledky ukázali napr. že čím väčšia perióda vzorkovania (dlhší horizont prognózy), tým menšia chyba predikcie. Medzi najvýznamnejšie rozhodovacie atribúty patria hmotnosť vozidla, rýchlosť vozidla, sklon cesty, smer vetra a rýchlosť vetra.

My chceme tento výskum posunúť smerom k podpore rozhodovania pri výbere vodiča na konkrétnu trasu. V súčasnosti je vo zvolenej firme prideľovanie vodičov vykonávané manuálne podľa toho, kde sa práve nachádza a či má vodič nárok na voľno. Podľa práce je odvolaný na cestu, tzv. „turnus“, ktorý trvá približne 20 – 25 dní [7]. Do rozhodovania pri prideľovaní vodičov ale určite vstupujú aj ďalšie dôležité charakteristiky, ktoré sa aktuálne nezohľadňujú.

Vďaka veľkému množstvu dostupných dát o jednotlivých vodičoch, ako napr. výkon vodiča, priemerná spotreba paliva, dodržiavanie maximálnej rýchlosti a tiež mnohé iné parametre, ktoré dokážu poskytnúť informácie napr. o štýle jazdy, priemernej rýchlosti a pod., môžeme získať celkový prehľad, pomocou ktorého je možné porovnávať, analyzovať a zostavovať modely a vykonávať experimenty pre rôznych vodičov a pre rôzne vozidlá [4].

2 Popis dát

K dispozícii sme mali dáta, ktoré boli získané prostredníctvom systému Danafleet Online – Volvo Truck Corporation. Firma tento systém využíva na komunikáciu s vozidlami. Vozidlá generujú informácie, ktoré sú prostredníctvom mobilnej telefónnej siete odosielané do systému. Z vozidiel sú ukladané do databázy a následne do výkazov. Dáta je možné z výkazov exportovať do súboru MS Excel za účelom ďalších analýz. K dispozícii sme mali dve dátové množiny od troch rôznych vozidiel. Prvá dátová množina bola získaná z výkazu hodnotenia spotreby paliva a reprezentovala jazdný štýl daného vozidla k určitému dátumu. Okrem toho sa v dátovom súbore nachádzali aj ďalšie numerické atribúty ako: dátum, celkový čas, celková vzdialenosť, celkové hodnotenie, priemerná rýchlosť, priemerná spotreba paliva (l/100 km), celkové splodiny CO₂, predvídanie, voľný dojazd, využitie motora, zaťaženie, prekročenie rýchlosti, prispôbenie rýchlosti, tempomat, voľnobeh... Druhý dátový súbor bol získaný zo zostavy sledovania a obsahoval atribúty ako: dátum, meno vodiča, stav paliva, prejdená vzdialenosť, miesto...

Predspracovanie dát spočívalo v generalizácii miesta na konkrétny štát, v ktorom sa vozidlo nachádzalo. Následne sme redukovali atribúty a z druhého dátového súboru

sme vybrali len dátum, meno vodiča a miesto. Pomocou atribútu dátum sme zlúčili obidva dátové súbory jedného vozidla. Takéto spojenie sme vykonali aj pre ďalšie dve vozidlá. Následne sme získali tri dátové vzorky, ktoré sme zlúčili do jednej. Po tomto zlúčení sme odstránili duplicitné hodnoty a silne korelované atribúty (korelácia vyššia ako 0.9) a vykonali diskretizáciu jednotlivých numerických atribútov (celkové hodnotenie, predvídanie, zaťaženie motora...), a to nasledovne:

- 0 – 59: Dobrý výkon
- 60 – 79: Priemer
- 80 – 100: Potenciál k zlepšeniu

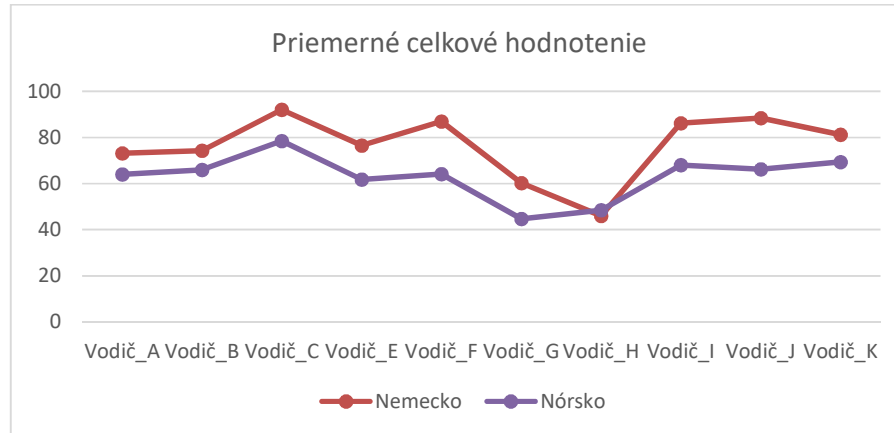
Po tejto úprave sme mohli začať s návrhom experimentov a modelovaním.

2.1 Návrh experimentov a modelovanie

Cieľom modelovania bolo zistiť, do akých krajín jazdia jednotliví vodiči najčastejšie a aké majú hodnotenie jazdného štýlu v danej krajine. Celkové hodnotenie predstavuje spôsob hodnotenia jazdného štýlu vodiča v určitom okamihu. Do celkového hodnotenia sa berú do úvahy aj atribúty ako Predvídanie, Prispôsobenie rýchlosti, Využitie motora a prevodovky a Úplné zastavenie. Tieto hodnoty systém berie do úvahy, na základe čoho dokáže vyjadriť percentuálne hodnotenie vodiča v danom okamihu.

Z jednoduchšej štatistickej analýzy sme zistili, že krajina ako je Nemecko a Nórsko bola najčastejšie navštevovaná všetkými desiatimi vodičmi. Z grafu je možné vidieť, že v Nemecku dosahovali vodiči lepšie hodnotenie ako v Nórsku. Je to aj z toho dôvodu, že v Nemecku vodiči jazdia prevažne po diaľnici, v Nórsku musia zase prekonávať výškové rozdiely a jazdia prevažne po horských priechodoch, čo má výrazný vplyv na výkon vozidla.

Z grafu taktiež môžeme vidieť ako jednotliví vodiči zvládajú jazdu v danej krajine. Môžeme napr. povedať, že Vodič_F má výrazne lepšie hodnotenie ako Vodič_H. Takáto analýza môže pomôcť majiteľovi firmy pri rozhodovaní akého vodiča priradiť na akú trasu.

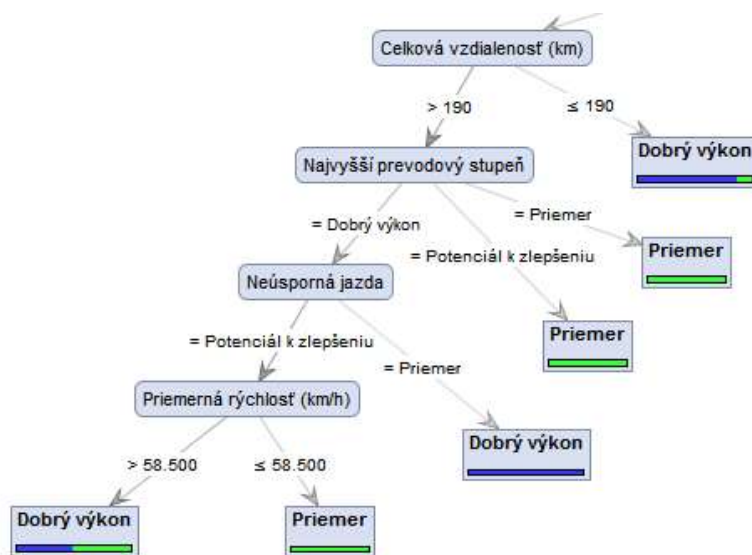


Obr. 1 Graf celkového hodnotenia jednotlivých vodičov.

Ďalší experiment mal skôr popisný charakter. Chceli sme zistiť kombináciu faktorov, ktoré majú kľúčový vplyv na výkon vodiča, pričom sme využili rozhodovacie stromy. Použili sme 20-násobnú krížovú validáciu a ako kritérium pre výber atribútu sme použili informačný zisk. Cieľovou premennou bolo celkové hodnotenie vozidla (Obr. 2), pričom modrá farba znamená, že vodič dosiahol v celkovom hodnotení jazdného štýlu dobrý výkon (hodnoty pred diskretizáciou 80 – 100), zelená farba znamená, že vodič dosiahol v celkovom hodnotení jazdného štýlu priemerný výkon (hodnoty pred diskretizáciou 60 – 79).

2.2 Vyhodnotenie rozhodovacieho stromu

Presnosť nami vytvoreného modelu je 83,52 %. Cieľom rozhodovacieho stromu bolo určiť kombináciu faktorov, ktoré majú kľúčový vplyv na výkon vodiča. Z výsledného rozhodovacieho stromu môžeme teda povedať, že podstatnými atribútmi sú Predvídanie, Priemerná spotreba paliva (l/100 km), Priemerná rýchlosť, Celková vzdialenosť (km) ako aj Najvyšší prevodový stupeň, Neúsporná jazda, Pomer brzdení/zastavenia. Tieto atribúty majú vplyv na spôsob hodnotenia jazdného štýlu vodiča. Analýza takýchto dát môže odhaliť oblasti, v ktorých je možné znížiť spotrebu paliva a poskytnúť vodičom tipy, v čom sa môžu zlepšiť.



Obr. 2 Rozhodovací strom (výsek).

3 Záver a budúca práca

V práci sme sa zamerali na analýzu existujúceho procesu priradzovania vodičov na jazdné trasy vo zvolenej logistickej firme. Pomocou jednoduchšej štatistickej analýzy sme poukázali na skutočnosť, ako jednotliví vodiči jazdia v jednej z dvoch najčastejšie navštevovaných krajín. Takáto analýza môže pomôcť majiteľovi vyriešiť rozhodnutie o správnom priradení vodiča na plánovanú dodaciu trasu vzhľadom k výkonu vodiča a k jeho jazdnému štýlu.

Cieľ ďalšieho výskumu bude zameraný na vytvorenie ako aj overenie modelu vodiča a modelu dodacej trasy. Na základe analýzy dát ako aj vykonaných rozhovorov s vodičmi a ďalšími relevantnými aktérmi procesu zostavíme model zahrňujúci všetky rozhodujúce faktory ovplyvňujúce výkon vodiča, resp. charakteristiky dodacej trasy. Následne vytvoríme a experimentálne overíme systém pre podporu rozhodovania o priradzovaní vodičov na dodacie trasy vo zvolenej logistickej firme. Meraním výkonnosti organizácie po a pred nasadením systému pre podporu rozhodovania zistíme do akej miery je tento systém prínosom pre zvolenú logistickú spoločnosť.

Pod'akovanie. Táto publikácia vznikla vďaka podpore Vedeckej grantovej agentúry MŠVVaŠ SR a SAV projekt č. 1/0493/16.

Literatúra

1. Berglund, M., Laarhoven, P., Sharman, G.: Third-Party Logistics: IS There a Future. In: The International Journal of Logistics Management, (2006), pp. 59-70, ISSN: 0957-4093
2. Congna, Q., Huifeng, Z.: Study on Application of Data Mining Technology to Modern Logistics Management Decision. In: International Forum on Information Technology and Applications, (2009), pp. 433-436, ISBN: 978-0-7695-3600-2
3. Gustin, C. M., Stank, T.P.: Computerization: Supporting Integration. In: International Journal of Physical Distribution & Logistics Management, (2006), pp. 11-16, ISSN: 0960-0035
4. Houghton, M. A.: Assigning Delivery Routes to Drivers under Variable Customer Demands. In: Transportation Research Part E: Logistics and Transportation Review: ScienceDirect, (2007) pp. 157-172, ISSN: 1366-5545.
5. Laxhammar, R., Gascón-Vallbona, A.: D4.3 Vehicle models for fuel consumption. In: Co-operative dynamic formation of platoons for safe and energy-optimized goods transportation. Deliverable, FP7 EC project Companion, (2015), 62 p.
6. Lin, Cheng-Chang, Lin J. S-J.: The Multistage Stochastic Integer Load Planning Problem. In: Transportation Research Part E: Logistics and Transportation Review: ScienceDirect, (2007), pp. 143-156, ISSN: 1366-5545.
7. Muchová, M: Big data analysis in selected logistics process. In: SCYR 2016: Proceedings from Conference. Košice: TU, (2016), pp. 55-57. ISBN 978-80-553-2566-8
8. Paul A., Saravanan, V.: Data Mining Analytics to Minimize Logistics Cost. In: International Journal of Advances in Science and Technology. (2011), pp. 89-107, ISSN: 2229-5216

Annotation:

Data analysis to improve specific business of processes logistics company.

Data mining is very important for modern logistics management, which helps to improve the right decisions, increase sales, reduce costs, and so on. In the context of these decisions, not only the correct information and knowledge are essential, but also a way how to use them effectively. The article deals with data analysis to improve decision-making in the selected logistics process - selection of drivers for planned delivery routes. The paper presents our initial steps in addressing individual subtasks using decision trees and simple statistical methods that we have applied to the data from the selected company. The obtained results are described in this paper.

Rozpoznanie podobnosti textov, programových kódov

Juraj Petrík, Daniela Chudá

Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava, Slovenská republika

{juraj.petrík, daniela.chuda}@stuba.sk

Abstrakt. Práca sa zaoberá rozpoznávaním podobnosti v kontexte plagiátorstva v programových kódach. Na zistenie, ako náročné je vytvorenie dokonalého plagiátu bol vykonaný experiment, ktorý ukázal, že vytvorenie takéhoto plagiátu je pomerne jednoduché a nevyžaduje hlbšie znalosti programovacieho jazyka. Na základe tohto experimentu bol navrhnutý nástroj PerfectPlaggie, ktorý je schopný automatickej tvorby programových klonov za účelom tvorby dátovej množiny na ďalšie testovanie. Ďalej je opísaný antiplagiátorský systém používaný na FIIT STU, ktorý je unikátny tým, že sa snaží o maximálne využitie štandardných Unixových filtrov.

Typ príspevku: Doktorandské sympóziu

Kľúčové slová: podobnosť, zdrojový kód, klon, dokonalý plagiát, PerfectPlaggie

1 Úvod

Každým dňom je ľudstvom vytvorené obrovské množstvo dát – nachádzame sa v ére takzvaných veľkých dát. Avšak veľké množstvo týchto dát sú dáta podobné, niekedy až rovnaké.

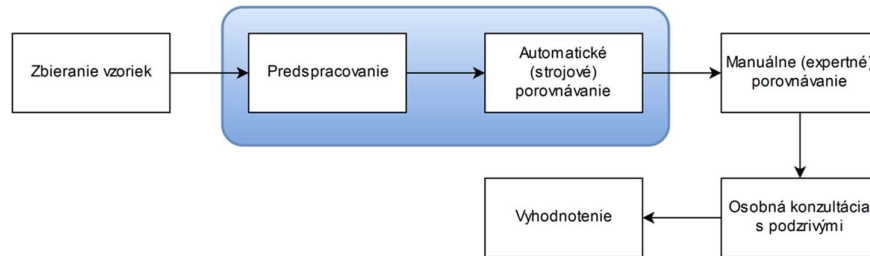
Podobnosť rozpoznávame napríklad za účelom personalizovaného odporúčania, identifikácie nevyžiadanej pošty, refaktoringu, detekcie plagiátov alebo aj za účelom odhaľovania škodlivého kódu (malvéru).

Detekcia plagiátorstva je otvorený a významný problém. Pretože s rozširovaním internetu a tým súvisiacej voľnej dostupnosti veľkého množstva dát ako sú texty alebo zdrojový kód sú ľudia pokúšaní zvoliť si „ľahšiu cestu“.

Plagiátorstvo je definované ako uvádzanie myšlienok alebo textov niekoho iného za svoje vlastné.¹ Štúdie [1,2] ukazujú, že problém plagiátorstva v akademickej sfére je ešte vážnejší, ako sme si doteraz mysleli.

Úloha automatickej rozpoznávanie podobnosti v celom procese určovania plagiátov je zvýraznená na obrázku 1.

¹ <http://www.merriam-webster.com/dictionary/plagiarized>



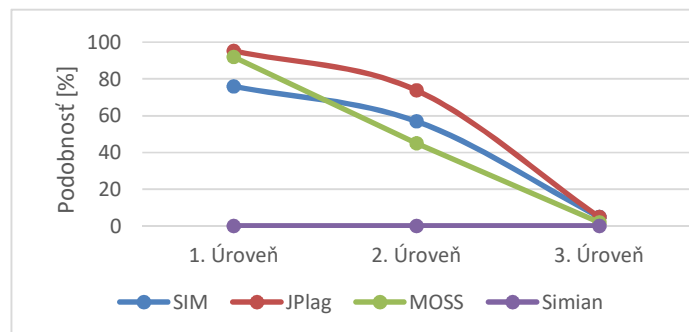
Obr. 1 Podobnosť v procese hľadania plagiátov

2 PerfectPlaggie

PerfectPlaggie je názov experimentu a následne aj názov navrhovaného nástroja. Experiment, ako aj nástroj sa zaoberá tvorbou dokonalého plagiátu. Pod dokonalým plagiátom rozumieme taký plagiát, ktorý nie sú schopné dostupné nástroje a metódy označiť za podozrivý.

2.1 Experiment manuálnej tvorby programového klonu

Pri tomto experimente bolo sledované aké veľké úsilie a aké množstvo vedomostí je potrebné na zakrytie plagiátorstva v programových kódoch. Za referenčné nástroje, ktoré vyhodnocovali mieru podobnosti boli vybrané nástroje SIM, JPlag [4], MOSS a Simian. Úpravy boli rozdelené do troch úrovní podľa zložitosti vykonania daných zmien.



Obr. 2 Nájdené percento podobnosti

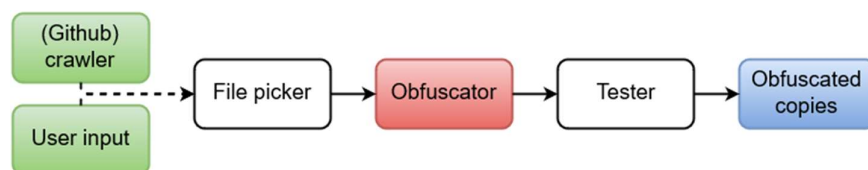
Obrázok 2 zobrazuje výsledky - nájdenú mieru podobnosti medzi originálom a upravenou verziou (klonom, plagiátom). Ako je možné vidieť, po úpravách tretej úrovne nie je ani jeden z testovaných nástrojov schopný odhalenia podobnosti. Podľa [3,5] majú

všetky dostupné nástroje a metódy problémy pri použití sofistikovaných typov zakrývania. Zmeny, ktoré boli vykonávané, boli vykonávané tak, aby simulovali vytváranie týchto zmien pomocou počítača, čo znamená, že tieto zmeny je možné vytvárať aj automatizovane.

2.2 Automatizovaná tvorba klonov

Manuálne vytváranie veľkých dátových množín, ktoré sú potrebné, je zdĺhavé a môže viesť ku chybám pri upravovaní zdrojového kódu. Výhodou takto vytvorenej dátovej množiny je to, že máme istotu, že dané dvojice sú klonmi a navyše máme aj informáciu ako tieto klony vznikli.

Obrázok 3 znázorňuje architektúru nástroja PerfectPlaggie, ktorý je navrhnutý na automatickú tvorbu programových klonov. Tento nástroj je navrhnutý tak, aby automaticky sťahoval projekty s otvoreným zdrojovým kódom z internetu, vyberal z nich tie zaujímavé (podľa rôznych metrik) a následne tieto súbory modifikoval (rôznymi typmi zakrývania). Výstupom tohto nástroja budú zmodifikované súbory, ktoré vznikli z originálneho súboru. Tieto súbory budú mať rovnakú funkčnosť ako originál.



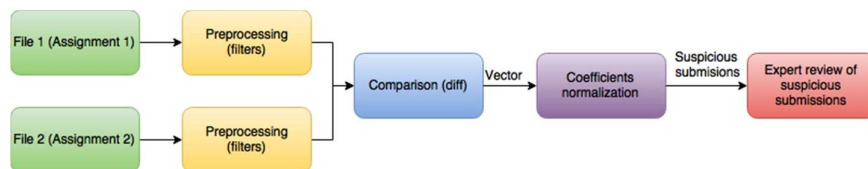
Obr. 3 Návrh architektúry nástroja PerfectPlaggie

Pri implementácii zakrývania sa berie aj ohľad na to, že podozrivé súbory môžu byť kontrolované ľudským expertom. A teda napríklad pri zmene názvov budeme vychádzať zo synonymického slovníka.

3 Identifikácia plagiátov na FIIT

Používaný systém na detekciu plagiátov na FIIT STU je odlišný od iných systémov tým, že sa snažil v maximálnej možnej miere využívať štandardné Unixové filtre. Tento nástroj sa skladá z dvoch častí – kontroly zadání a normalizácií koeficientov.

Časť kontrola zadání porovnáva dvojice súborov. Každá kontrola sa skladá z viacerých úrovní. Prvá úroveň je iba textové porovnanie súborov (diff), druhá je textové porovnanie súborov, avšak ignorujú sa už prázdné riadky a biele znaky. Tretia úroveň je druhá úroveň spolu s tým, že všetky alfanumerické znaky sú nahradené znakom X – čím získame „štruktúru“ programu.



Obr. 4 Proces vyhodnocovania podobnosti pre jednu úroveň

Štvrtá úroveň je druhá úroveň s odstránenými komentármi, navyše každé slovo je na samostatnom riadku. Piata úroveň má navyše oproti štvrtej úrovni usporiadanie riadkov podľa abecedy. Výstupom tejto časti je šesťzložkový vektor, ktorý obsahuje informácie o počte riadkov spracovaných súborov, počte potrebných zmien, aby sme mali identické súbory a priemernej veľkosti zmeny.

Časť normalizácie koeficientov upravuje vektor z predchádzajúcej časti a na základe heuristicky zistených hodnôt určuje podozrivosť dvojice.

Obrázok 4 zobrazuje porovnávanie jednej dvojice výsledky z jednej časti putujú do ďalšej časti (rúrovité spracovanie). Konečným výstupom je zoznam podozrivých dvojíc.

4 Záver

Experiment dokonalého plagiátu ukázal, že vytvorenie dokonalého plagiátu nie je zložité a že úsilie vložené do tvorby takéhoto plagiátu je nižšie, ako úsilie, ktoré by bolo potrebné na samostatné naprogramovanie riešenia zadania. Na základe tohto experimentu vznikol návrh nástroja PerfectPlaggie, ktorý slúži na vytváranie veľkých dátových množín, ktoré budú použité na ďalšie testovanie nových nástrojov a metód.

Antiplagiátorský systém používaný na FIIT STU má zaujímavú architektúru, avšak je potrebné ešte ďalšie preskúmanie a otestovanie tohto systému, aby mohli byť navrhnuté zmeny, ktoré by viedli k ešte lepším výsledkom.

Pod'akovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektu Prispôsobovanie prístupu k informačným a vedomostným artefaktom založené na interakciách a kolaborácii v prostredí webu, Vedecká grantová agentúra MŠVVaŠ SR a SAV, grant No.. VG 1/0646/15.

Literatúra

1. ARWIN, C., TAHAGHOGHI, S.M.M. Plagiarism Detection across Programming Languages. In *Twenty-Ninth Australasian Computer Science Conference (ACSC2006)*. 2003. s. 10.
2. CHUDA, D. et al. The issue of (software) plagiarism: A student view. In *IEEE Transactions on Education*. 2012. Vol. 55, no. 1, s. 22–28.
3. POTTHAST, M. et al. Overview of the 6th International Competition on Plagiarism Detection. In *Notebook for PAN at CLEF 2014*. 2014. s. 845–876.

4. PRECHELT, L. et al. Finding Plagiarisms among a Set of Programs with JPlag. In Journal Of Universal Computer Science. 2002. Vol. 8, no. 11, pp. 1016–1038.
5. TAHIR ALI, A.M. EL et al. Overview and comparison of plagiarism detection tools. In *CEUR Workshop Proceedings* . 2011. Vol. 706, s. 161–172.

Annotation:

Recognizing similarity of texts, programming codes.

The work deals with the similarity in context of plagiarism in source codes. To find out how difficult it is to create perfect plagiarism we have done an experiment, that showed that creation of such plagiarism is quite simple and does not require much knowledge of programming language. PerfectPlaggie tool was designed according to this experiment. This tool is capable of creating software clones in order to automatically create large datasets. Also there is described plagiarism detection system, which is used at FIIT STU. This system is utilizing standard Unix filters to detect plagiarism.

Automatické generovanie prediktorov a ich využitie pri dolovaní pravidiel

Michal Puheim, Kristína Machová, Ján Paralič

Fakulta elektrotechniky a informatiky
Katedra kybernetiky a umelej inteligencie
Letná 9, 042 00 Košice, Slovenská republika

{michal.puheim, kristina.machova, jan.paralic}@tuke.sk

Abstrakt. V tomto príspevku prezentujeme systém pre dolovanie pravidiel nasaďený vo forme cloudovej služby, ktorý je určený pre analýzu veľkých dát. Cieľom systému je analýza veľkého množstva údajov o rôznych udalostiach s využitím prostriedkov dátovej agregácie, zhľukovania, klasifikácie a predikcie. Systém pozostáva z dvoch komponentov implementovaných ako sieťové služby. Generátor prediktorov zabezpečuje zmysluplný spôsob agregácie veľkého množstva údajov a Extraktor pravidiel správania sa venuje analýze týchto agregácií. Výsledkami systému sú predikčné pravidlá použiteľné pri podpore rozhodovania v oblastiach manažmentu, marketingu, segmentácie zákazníkov, klasifikácie, predikcie správania, atď.

Typ príspevku: Doktorandské sympóziu

Kľúčové slová: agregácia dát, prediktory, dolovanie pravidiel, cloudová služba

1 Úvod

Technologický pokrok v posledných desaťročiach vedie k nárastu potreby spracúvať veľké množstvo dát v rôznych oblastiach praxe. Veľké dáta [1] ovplyvňujú náš každodenný život aplikáciami, ktoré zahŕňajú zdravotné systémy, sociálne systémy, veľké vedecké experimenty, úložiská dát, logistiku a donáškové služby a pod. Hlavným problémom veľkých dát je ich množstvo, rôznorodá štruktúra a z toho vyplývajúce problémy so spracovaním v reálnom čase. Množstvo spracovávaných dát sa v súčasnosti dostáva na úroveň exaškály, kde jediný analytický systém potrebuje spracúvať vyše 10^{18} výpočtov za sekundu [2]. Je zrejmé, že tento vývoj vyžaduje nové systémové architektúry pre akvizíciu, transfer, ukladanie a spracovanie dát. Rastúci počet aplikácií začína využívať vrstvené systémové architektúry a cloudovú infraštruktúru s cieľom zvládnuť požiadavky na spracovanie tohto množstva dát [2].

1.1 Popis prebiehajúceho výskumu

Pri implementácii projektu [3] sa zameriavame na aplikačný výskum a vývoj softvérových riešení umožňujúcich efektívne riešenie rôznych problémov v oblasti analýzy veľkých dát, vrátane predaja, marketingu, personalizácie a odporúčaní, manažmentu rizika, optimalizácie výrobných procesov, skvalitňovania zdravotnej starostlivosti a pod [3][4]. Jedným z výstupov projektu bude softvérová platforma pre analýzu a optimalizáciu procesov, poskytovaná vo forme softvéru ako služby.

Náš výskumný tím je rozdelený na niekoľko skupín, ktoré sa zaoberajú mimo iného: analýzou veľkých textových dát, aspektovo-založenou analýzou sentimentu v dokumentoch, analýzou zdravotníckych dát, dolovaním pravidiel v procesoch a udalostiach a personalizovanými odporúčaniami.

V rámci poslednej skupiny sa venujeme analýze správania zákazníkov založenej na spracovaní procesných záznamov (logov) a udalostí (napr. návšteva stránky, zobrazenie položky v e-shope, účasť v emailovej kampani a pod.). Cieľom analýzy je získanie znalostí vo forme klasifikačných a predikčných pravidiel, segmentácie zákazníkov na základe podobného správania a hľadanie charakteristických vzorcov správania.

1.2 Definícia pojmu prediktor

Pri analýze logov predpokladáme semi-štruktúrovanú podobu dát vo forme tabuľky udalostí a odpovedajúcej tabuľky entít (napr. zákazníkov) zodpovedných za tieto udalosti. Každý riadok v tabuľkách reprezentuje jednu entitu alebo udalosť, pričom každej entite môže prislúchať viacero udalostí. Zároveň predpokladáme neustály nárast množstva údajov v tabuľke udalostí, ktorý znemožňuje priamu analýzu týchto údajov. Z tohto dôvodu údaje o udalostiach najprv agregujeme do novo vytvorených atribútov v tabuľke entít (Obr. 1) a až tie následne analyzujeme.

| UDALOSTI | | | | | | | | | |
|--------------------------|-------------------|-------------------|---------------------|------------------|----------------|-----------------|----------------|------------------------|--|
| customer_id | purchase_category | purchase_category | purchase_product_id | purchase_product | purchase_count | purchase_profit | purchase_price | purchase_selling_price | |
| 53039a0c25030f78d9d4dbfd | 1017 | Shoes | 122941 | AUTHORI | 1 | 5.98 | 7.01 | 12.99 | |
| 53039a0c25030f78d9d4dc01 | 1017 | Shoes | 112049 | ADIDAS- | 1 | 5.85 | 29.34 | 39.99 | |
| 53039a0c25030f78d9d4dc05 | 1001 | accessory | 108737 | EXISPOR | 1 | 1.55 | 1.41 | 2.99 | |
| 53039a0c25030f78d9d4dc05 | 1012 | football ITG | 113814 | NIKE- | 1 | 6.84 | 28.15 | 34.99 | |
| | | | | | 1 | 3.75 | 6.04 | 9.79 | |
| | | | | | 1 | 8.97 | 7.02 | 15.99 | |
| | | | | | 1 | 9.14 | 30.85 | 39.99 | |
| | | | | | 1 | 0.87 | 0.59 | 1.46 | |
| | | | | | 1 | 1.39 | 0.39 | 1.95 | |
| | | | | | 1 | 1.39 | 0.39 | 1.95 | |

| ENTITY | | | | | | | |
|--------------------------|-----------------------|----------------------|-------------------|-----------------------|--------------------|---------------------|--------------------|
| customer_id | properties_first_name | properties_last_name | properties_gender | properties_birth_year | purchase_count_sum | purchase_profit_avg | purchase_price_avg |
| 53039a0c25030f78d9d4dbfd | Katarina | Pe | ova | female | 960 | 1 | 5.98 |
| 53039a0c25030f78d9d4dc01 | Božena | Re | ová | female | 1 | 10.65 | |
| 53039a0c25030f78d9d4dc05 | Mariana | Si | ia | female | 1972 | 15 | 4.28 |
| 53039a0c25030f78d9d4dc09 | Stanislav | Bi | ak | male | 3 | 8.15 | |
| 53039a0c25030f78d9d4dc0a | Rudolf | Ac | ov | male | 1972 | 1 | 7.25 |
| 53039a0c25030f78d9d4dc0e | František | Ha | n | | | | |

Agregované atribúty / prediktory

Obr. 1 Princíp generovania agregovaných atribútov – prediktorov.

Keďže funkcií umožňujúcich túto agregáciu je teoreticky nekonečné množstvo, našim cieľom je identifikácia iba tých agregácií, ktoré majú vysokú koreláciu s iným (cieľovým) atribútom v tabuľke entít, a teda je možné ich využiť pri generovaní predikčných pravidiel. Takéto korelované agregované atribúty nazývame *prediktory* [5].

2 Predstavenie systému

Navrhnutý analytický systém (bližšie špecifikovaný v konferenčnom príspevku [6]) pozostáva z dvoch hlavných komponentov realizovaných vo forme nezávislých sieťových služieb. Prvým je *Automatický generátor prediktorov* (APG – *Automatic Predictor Generator*), ktorý pripravuje prediktory. Tie sú následne spracované v druhom komponente, pracovne nazvanom *Automatický extraktor pravidiel správania* (ABRE – *Automatic Behavior Rule Extractor*), ktorý nad tabuľkou entít generuje rozhodovacie stromy a z nich extrahuje najvýznamnejšie pravidlá.

Slovný popis algoritmu APG je nasledovný: 1) Služba akceptuje najmenej jednu tabuľku entít a jednu alebo viacero tabuliek udalostí. 2) V týchto tabuľkách deteguje dátový typ všetkých atribútov. 3) Súčasne s tabuľkou je službe určený cieľový atribút v tabuľke entít, voči ktorému sú následne generované nové prediktory. 4) Pri generovaní sú použité funkcie *počtu*, *súčtu*, *priemeru*, *maxima*, *minima* a *rozptylu* pre numerické atribúty a funkcie *počtu*, *počtu unikátov*, *najčastejšieho výskytu* a *názvu najfrekvencovanejšieho atribútu* pre nominálne atribúty. 5) Novo vygenerované atribúty sú následne filtrované s využitím vlastnej implementácie *Hierarchického aglomeratívneho zhľukovania* (HAC) [7]. Metrikou pre zhľukovanie je *Pearsonov korelačný koeficient* pre numerické atribúty a *Chi-kvadrát test nezávislosti* pre nominálne atribúty. Výsledkom zhľukovania sú skupiny prediktorov s nízkou vzájomnou koreláciou, pričom prediktory v každej skupine sú usporiadané podľa klesajúcej korelácie voči cieľovému atribútu. 6) APG nakoniec aktualizuje tabuľku entít doplnením vopred definovaného počtu najlepších prediktorov z jednotlivých zhľukov.

Aktualizovaná tabuľka entít môže byť následne analyzovaná pomocou ABRE, ktorého algoritmus je nasledovný: 1) Služba akceptuje tabuľku entít a požadovaný cieľový atribút. 2) S ohľadom na cieľový atribút je vygenerovaný rozhodovací strom s použitím štandardnej procedúry *Rekurzívneho delenia* [8], kde ako selekčné kritérium je použitý *Gini Index* [9]. 3) Nakoniec sú z rozhodovacieho stromu extrahované [10] pravidlá s najväčším pokrytím a istotou, ktoré sú zároveň výsledkom celkovej analýzy a predstavujú nové znalosti použiteľné pre interpretáciu budúceho správania entít.

2.1 Implementačné poznámky k modulom APG a ABRE

Modul služby APG je implementovaný pomocou jazyka Python s využitím balíkov SciPy a NumPy pre výpočet korelačných koeficientov. Generátor agregovaných atribútov je našej vlastnej implementácie, rovnako ako HAC filter [7]. Komunikačné rozhranie služby je založené na mikro-serveri Flask. Služba umožňuje spracovanie správ vo formáte JSON (a v obmedzenej miere aj CSV súborov). Modul ABRE je v súčasnosti

implementovaný ako proces v analytickom nástroji Rapid Miner Studio a bude exportovaný do podoby webovej služby s využitím prostriedkov Rapid Miner Servera. Obe služby budú implementované s použitím kontajnerovej technológie Docker v kvázi-separovaných virtuálnych prostrediach. Kontajnery pre modul APG a pre Rapid Miner Server sú dokončené a preliminárne otestované. Implementácia ABRE je naplánovaná v najbližšom období.

3 Predbežné výsledky a diskusia

Vzhľadom k stále prebiehajúcej implementácii systému a podpornej cloudovej infraštruktúry [3], nebolo doposiaľ možné jeho celkové testovanie a overeniu bola podrobená iba základná funkčnosť jednotlivých modulov. Na tento účel bola použitá databáza historických údajov o zákazníkoch (entitách) a príslušných nákupných udalostiach, pozostávajúca zo 7 461 záznamov zákazníkov a 10 002 záznamov udalostí. Tabuľka zákazníkov obsahovala dva aplikovateľné atribúty, z ktorých jeden bol zvolený ako cieľový atribút výslednej analýzy. Tabuľka udalostí obsahovala 8 aplikovateľných atribútov s celkovým počtom 80 016 hodnôt. Keďže dva atribúty boli nominálne a šesť numerických, algoritmus APG vygeneroval $2 \cdot 4 + 6 \cdot 6 = 44$ unikátnych agregovaných atribútov. Tieto boli následne filtrované s cieľom získať požadované tri najlepšie prediktory. Celé spracovanie prebehlo v rámci modulu na jednom jadre 2GHz CPU v čase nižšom než 5 sekúnd. Výsledné prediktory (spolu so zvyšným nepoužitým atribútom tabuľky zákazníkov) boli použité v module ABRE na vygenerovanie 17 pravidiel v čase približne 3 sekundy (rovnako na jednom jadre 2GHz CPU). Najlepšie pravidlo pokrylo 6 581 príkladov, z ktorých správne klasifikovalo 3600. Druhé najlepšie pravidlo pokrylo 496 príkladov a správne klasifikovalo 322.

Tieto výsledky potvrdzujú základnú funkčnosť navrhnutého konceptu cloudovej služby najmä z pohľadu výkonu. Samozrejme, cieľom do budúcnosti je zlepšiť výsledky analýzy, čo je možné dosiahnuť viacerými spôsobmi. V prípade modulu APG je možné okrem súčasných agregáčnych funkcií uvažovať aj o ďalších spôsoboch generovania nových prediktorov. V prípade modulu ABRE je potrebné doladiť parametre použité pre generovanie rozhodovacieho stromu, resp. generovať tieto pravidlá priamo. Hlavným cieľom je však dokončenie implementácie rozhraní medzi cloudovými službami a ich nasadenie na sprevádzkovanú cloudovú infraštruktúru [3].

4 Záver

V tomto článku sme predstavili základný smer výskumného projektu [3] realizovaného na našom pracovisku a popísali sme jeden z jeho výstupov [6]. S ohľadom na uľahčenie popisu sme špecifikovali základnú definíciu pojmu *prediktor*. Následne sme popísali algoritmus služby APG, ktorá je schopná agregovať relatívne veľké množstvá dynamicky generovaných údajov z procesných logov, a algoritmus služby ABRE, ktorá umožňuje agregované údaje zmysluplne analyzovať. Nakoniec sme uviedli predbežné výsledky testovania modulov reprezentujúcich jednotlivé služby.

Pod'akovanie: Ďakujeme podpore projektu KEGA, č. 014TUKE-4/2015 – "Digitalizácia, virtualizácia a testovanie malého prúdového motora a jeho častí pomocou stendov pre potreby modernej aplikovanej výuky" (50%) a projektu Univerzitný vedecký park TECHNICOM pre inovačné aplikácie s podporou znalostných technológií, ITMS: 26220220182, podporeného z Operačného programu „Výskum a vývoj“ financovaného z Európskeho fondu regionálneho rozvoja. "Podporujeme výskumné aktivity na Slovensku/Projekt je spolufinancovaný zo zdrojov EÚ" (50%).

Literatúra

1. Hu, H., Wen, Y., Chua, T.-S., Li, X.: "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," Access, IEEE, vol. 2, pp.652-687, 2014.
2. Branch, R., et al.: "Cloud Computing and Big Data: A Review of Current Service Models and Hardware Perspectives," Journal of Software Engineering and Applications, vol.7, no.8, pp. 686-693, 2014.
3. Paralič, J., et al.: "IT tools and services for analyses of different types of processes," In: University Science Park TECHNICOM, Košice, 2014, pp. 16.
4. Paralič, J., et al.: "IT tools and services for analyses of different types of processes," In: Modely fungovania vedeckých parkov a výskumných centier : skúsenosti a príležitosti pre Slovensko, Košice, 2015, pp. 53.
5. Paralič, J., Kováč, J.: "Automatické generovanie prediktorov zo semi-štruktúrovaných dát – Definícia projektovej úlohy", nepublikované, Máj 2014.
6. Puheim, M., Paralič, J., Madarász, L.: "Automatic predictor generator & behaviour rule extractor – A system proposal," In: Computational Intelligence and Informatics (CINTI), 2015 16th IEEE International Symposium on, Budapest, 2015, pp. 155-159.
7. Zepeda-Mendoza, M. L., Resendis-Antonio, O.: "Hierarchical Agglomerative Clustering," Encyclopedia of Systems Biology, Springer New York, 2013, pp. 886-887.
8. Friedman, J. H.: "A Recursive Partitioning Decision Rule for Nonparametric Classification", IEEE Transactions on Computers, vol.26, no. 4, pp. 404-408, April 1977.
9. Lerman, R. I., Yitzhaki, S.: "A Note on the Calculation and Interpretation of the Gini Index," Economics Letters, vol. 15, no. 3-4, pp. 363-368, 1984.
10. Quinlan, J. R.: "Generating Production Rules from Decision Trees." In: IJCAI. vol. 87. 1987.

Annotation:

Automatic Generation of Predictors Used for Rule-Mining

In this paper we present a proposal for a data-mining system deployed as a cloud service which is supposed to be used for a big data analysis. The main purpose of the system is the analysis of a vast number of event logs using means of data aggregation, clustering, classification and prediction. The system is composed of two components implemented as software services. The Automatic Predictor Generator is supposed to provide a meaningful way to aggregate large amounts of data and the Automatic Behavior Rule Extractor deals with proper analysis of these aggregations. Results of the system are the prediction rules usable for support of decision-making and in areas such as management, marketing, customer segmentation, classification, behavior prediction etc.

Structural Features Extraction from Text using Applicative Supercombinator Form

Michal Sičák, Ján Kollár

Department of Computers and Informatics
Technical University of Košice
Letná 9, 040 01 Košice, Slovak Republic

{michal.sicak, jan.kollar}@tuke.sk

Abstract. Grammar representation offers useful features that can be used in other aspects of computing than the standard language interpretation. One of such aspects that is addressed in this paper, is representation of any meaningful written text in a single, non-redundant form. Such a form stores each distinct word separately, thus reduces the entire size of a document. Another size reducing feature of the grammar form is its ability to abstract structure away from its content. Therefore by using lambda calculus application principle, we can create a supercombinator form of text substructures. This form, when applied on the arguments which are words themselves, produces the original text back. We show that this form offers reduction in total amount of language elements. We also show that supercombinators represent reusable language elements that can be used across analysed texts.

Contribution type: PhD Symposium

Keywords: text structure analysis, supercombinator form, structure extraction

1 Introduction

Grammars are widely used formalism generally used as a tool for language representation. But as Klint, Lämmel and Verhoef pointed out in [4], we can use them for other purposes as well, for example data compression, structure representation or feature extraction. The issue addressed in this paper revolves around representation of written natural language text in a compact, non-redundant form. We can compress text with the use of context free grammars, as Nevill-Manning and Witten have done in [7] where they've used Sequitur algorithm. But this form leaves us with many repetitive structures and symbols. We aim to reduce the total number of used elements.

This work correlates with the field of formal grammar inference [1, 3, 10] and natural language induction [2, 8]. We add upon those inferred grammar and store them in a non-redundant applicative supercombinator form. This form was devised in our previous work [5, 6], where we used only regular grammars as a basis for our experiments.

Recently, we extended this process to the realm of context free grammars, thus enabling the processing of inferred or induced grammars.

The main contributions of this paper are:

- We briefly present supercombinator set constructing algorithm that is capable to store any context free grammar in a non-redundant applicative form. This is the topic of Section 2.
- We present the results in the Section 3, where we compare the number of grammar elements of processed natural language text obtained from Sequitur algorithm and our compressed supercombinator form of that Sequitur grammar. We show that the amount of elements drops significantly, since our form is non-redundant. The original text can still be reconstructed from our form by simple function application.

2 Supercombinator Form and its Construction

In this section, we show how we can represent regular grammar in a form of enriched lambda calculus, which is a basis of our supercombinator form. Such enriched lambda calculus is extended with the meta-operations of processed grammar.

Let's consider a simple regular expression (1). It generates either a sequence $a b$ or zero to n repetitions of a symbol c .

$$a b \mid (c)^* \quad (2)$$

We see that expression (1) contains three meta-operations: concatenation, alternation and Kleene star closure. So in this case, extended lambda calculus not only contains standard variable, lambda application and lambda abstraction operations but also expressions meta-operations as well. In the Table 1 we see the complete set of supercombinators constructed from the expression (1). We see the use of by meta-operations, as concatenation is represented by symbol $+$, alternative with symbol \mid and Kleene star as usual $()^*$. The second column contains possible arguments for each supercombinator. As the first supercombinator is unary, the comma separating arguments means that only one of them may be used. Arguments without any comma represent a sequence of arguments. The main (or top) supercombinator is L^3 , by which application we obtain original expression (1) back. Other supercombinators represent structure. We may notice that each supercombinator is constructed just with one meta-operation. We can decompose any regular grammar that way.

Tab. 1 Supercombinator form of expression (1).

| Supercombinator | Arguments |
|---|-----------|
| $L^0 = \lambda x_0. x_0$ | a, b, c |
| $L^1 = \lambda x_0. \lambda x_1. L^0 x_0 + L^0 x_1$ | ab |
| $L^2 = \lambda x_0. (L^0 x_0)^*$ | c |
| $L^3 = \lambda x_0. \lambda x_1. \lambda x_2. L^1 x_0 x_1 \mid L^2 x_2$ | abc |

2.1 Context free extension

What about CFGs? As we pointed out in [9], CFG non-terminals may be viewed as higher order jumps into another expression. This is incorporated in our supercombinator obtaining algorithm. Each rule of CFG is represented by its own top supercombinator. And each call of non-terminal inside of a rule body is therefore just a call of that supercombinator with its own arguments. Therefore it is possible to construct supercombinator form of any CFG, where the starting rule is represented as the top supercombinator. This supercombinator has as its arguments all possible terminal symbols, represented as a non-redundant list. The second non-redundancy effect is achieved by reusing supercombinators.

In the Table 1 we see different supercombinators. But they all represent some specific structure. The L^0 supercombinator represents ID function, and is being used three times for three different arguments in the case of expression (1). But the larger grammar is, the greater number of supercombinators is reused. For example, the L^1 supercombinator represents a sequence of any two distinct symbols. It is reasonable to believe that such supercombinator is rather heavily used in larger grammars. In the next chapter, we confirm this reasoning with an experiment.

3 Experimental Results

In order for our supercombinator algorithm to work, we need a grammar. Written text is a plain sequence of words. We would obtain only two supercombinators from it (one for Id function and the second would be a long Top supercombinator representing a sequence of n symbols). We can however construct simple CFG with the use of Sequitur algorithm [7]. Such CFG generates only the original text; it does not offer any generalisation. But for this experiment it is adequate, since we want to show the reduction of elements. We are going to use a book sample obtained from the King James Bible. We use the entire New Testament as our sample.

Table 2 shows the amount of Sequitur rules of certain arity (i.e. how many terminals and non-terminals a rule has) compared to operation arity of supercombinators (How many supercombinator calls act as operands of lambda calculus n -ary meta-operation, in this case only concatenation. For example both L^1 and L^3 from Table 1 are binary).

Tab. 2 Sequitur rules compared to Supercombinators, part 1.

| Arity | 0 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|------------|----|-------|-----|-----|----|----|----|----|-------|
| Sequitur | 1 | 13782 | 714 | 242 | 76 | 52 | 27 | 19 | 8 |
| Super-com. | 1 | 286 | 188 | 142 | 65 | 51 | 27 | 19 | 8 |
| Arity | 10 | 11 | 12 | 13 | 14 | 16 | 17 | 20 | Total |
| Sequitur | 6 | 4 | 2 | 1 | 3 | 2 | 1 | 1 | 14942 |
| Super-com. | 6 | 4 | 2 | 1 | 3 | 2 | 1 | 1 | 808 |

A rule with certain arity is always transformed to a supercombinator of the same operational arity. We see that the greatest reduction occurs in the lower arities, while going up from the arity of 7, there is no reduction at all. This is normal, since large arity rules are rather rare and they differ from each other in their structure.

There is also one supercombinator (and rule) not shown in the tables. It is the top one (as well as the starting rule) with the arity of 86461. It is accounted for however in the total result in the Table 2. Total count shows significant reduction of elements. Remember that any distinct symbol (in this case a word) is stored in our form only once, thus the element reduction is indeed substantial.

4 Conclusions

We have rather briefly shown a way to represent CFG in a non-redundant applicative supercombinator form. Such form offers significant element reduction of Sequitur grammars obtained from text. We show this on a sample obtained from The King James Bible. We have generated a Sequitur grammar from the sample and then processed it with our algorithm. We show that the reduction of used elements is indeed rather significant, where 14942 grammar rules are transformed into 808 supercombinators.

Literature

1. De La Higuera, C.: A bibliographical study of grammatical inference. *Pattern recognition*, (2005), vol. 38, no. 9, pp. 1332–1348.
2. Edelman, S.: On the nature of minds, or: truth and consequences, *Journal of Experimental & Theoretical Artificial Intelligence*, (2008), vol. 20, no. 3, pp. 181–196.
3. Gold, E.M.: Language identification in the limit, *Information and control*, (1967), vol. 10, no. 5, p. 447–474.
4. Klint, P., Lämmel, R., Verhoef, C.: Toward an engineering discipline for grammarware. *ACM Trans. Softw. Eng. Methodol.*, (2005), vol. 14, no. 3, pp. 331–380.
5. Kollár, J., Spišiak, M., Sičák, M.: Abstract language of the machine mind, *Acta Electrotechnica et Informatica*, (2015), vol. 15, no. 3, pp. 24–31.
6. Kollár, J., Sičák, M., Spišiak, M.: Towards machine mind evolution. In: *Computer Science and Information Systems (FedCSIS)*, IEEE, (2015), pp. 985–990.
7. Nevill-Manning, C.G., Witten, I.H.: Identifying hierarchical structure in sequences: A linear-time algorithm. *J. Artif. Intell. Res.(JAIR)*, (1997), vol. 7, pp. 67–82.
8. Onnis, L., Waterfall, H.R., Edelman, S.: Learn locally, act globally: Learning language from variation set cues, *Cognition*, (2008), vol. 109, no. 3, pp. 423–430.
9. Sičák, M.: Higher order regular expressions. In: *Engineering of Modern Electric Systems (EMES)*, IEEE, (2015).
10. Stevenson, A., Cordy, J.R.: Grammatical inference in software engineering: an overview of the state of the art, In: *Software Language Engineering*. Springer, (2013), pp. 204–223.

Hierarchické modelovanie témy nad prúdmi dát zo sociálnych sietí s využitím formálnej konceptovej analýzy

Miroslav Smatana, Peter Butka

Fakulta elektrotechniky a informatiky
Technická univerzita v Košiciach
Letná 9, 042 00 Košice, Slovenská republika

{miroslav.smatana, peter.butka}@tuke.sk

Abstrakt. Jednou z možných analýz neštruktúrovaných textov je modelovanie témy, ktoré sa snaží odkrývať skryté tematické štruktúry v týchto textoch. Modelovanie témy môže byť užitočné hlavne v kontexte sociálnych sietí, kde môže slúžiť pre analýzu v čase krízových situácií, zavedení nového produktu na trh, atď. V súčasnosti vzniklo niekoľko modifikácií klasických prístupov, ktoré vytvárajú hierarchiu tém. Tie ponúkajú často krát podrobnejšiu analýzu ako klasické prístupy. Cieľom prezentovaného článku je preto predstaviť niekoľko možných prístupov hierarchického modelovania témy založených na využití formálnej konceptovej analýzy, ktorá slúži na analýzu objekt-atribútových modelov. Článok taktiež ponúka experimentálne overenie jedného z prístupov na príspevkoch zo sociálnej siete Twitter.

Typ príspevku: Doktorandské sympóziu

Kľúčové slová: modelovanie témy, sociálne siete, formálna konceptová analýza, prúdy dát

1 Úvod

Sociálne siete sa stali v posledných rokoch jedným z najvýznamnejších komunikačných prostriedkov dnešnej doby. Denne sa na nich produkuje obrovské množstvo príspevkov napríklad na sociálnej sieti Twitter¹ je denne publikovaných okolo 340 miliónov príspevkov. Tieto príspevky často krát odrážajú názory a postoje používateľov na rozličné produkty, osoby, udalosti a pod.

Dáta zo sociálnych sietí si získavajú svoje miesto najmä v oblasti marketingu, kde ich správne pochopenie a reprezentácia môžu spoločnosti priniesť konkurenčnú výhodu. Môžu byť využité napríklad pri krízovej analýze, médiami, zavedení nového produktu na trh a pod.

¹ <https://twitter.com/>

Ako je vidieť analýza dát zo sociálnych sietí má širokú oblasť využitia. Avšak pri ich spracovaní narážame na niekoľko problémov z ktorých najhlavnejším je kvantita publikovaných príspevkov, kde manuálna analýza takého množstva dát by bola časovo veľmi náročná. Preto je potrebné túto činnosť zautomatizovať. Jedno z možností je použitie metód modelovania témy, ktoré nám ukázalo nový spôsob sumarizácie, vyhľadávania a prehľadávania textov. Základnou myšlienkou modelovania témy je odkrývanie skrytých tematických štruktúr medzi vstupnými textami.

Z toho dôvodu sa budeme v tomto článku venovať experimentálnym prístupom hierarchického modelovania témy s využitím formálnej konceptovej analýzy.

2 Formálna konceptová analýza

Formálna konceptová analýza (FCA) [1] patrí medzi metódy analýzy dát, ktorej popularita vzrástla najmä v posledných rokoch. Svoje využitie nachádza v mnohých oblastiach ako dolovanie v dátach, navracaní informácií, dolovanie asociačných pravidiel atď.

FCA sa dá využiť podobne ako techniky hierarchického zhľukovania, výsledkom tejto analýzy sú nájdené súvislosti v dátach tzv. konceptový zväz. Konceptový zväz reprezentuje kolekciu formálnych konceptov (typ hierarchie konceptov), ktoré sú hierarchicky zoradené.

Existuje niekoľko metód na budovanie konceptových zväzov niektoré z nich sú prezentované a porovnané v práci [2]. Ďalšou metódou využívanou aj v našej práci je zovšeobecnený jednostranne fuzzy konceptový zväz (GOSCL). GOSCL je model formálnej konceptovej analýzy, ktorý na generovanie konceptového zväzu využíva jednosmernú fuzzifikáciu. Výhodou tohto modelu je, že dokáže generovať koncepty z objektov pozostávajúcich z rôznych typov atribútov (nominálne, ordinálne, numerické a iné.) a taktiež patrí medzi inkrementálne algoritmy tj. je ho možné využiť na spracovanie prúdov dát. Viac informácií o GOSCL je možné nájsť v práci [3].

3 Modelovanie témy

V posledných rokoch bolo prezentovaných niekoľko prístupov k modelovaniu tém. Jedným z najpopulárnejších je Latentná Dirichletova Alokácia (LDA) [4]. Z LDA bolo vytvorených niekoľko jej rozšírení [5, 6]. Taktiež bolo predstavených niekoľko odlišných prístupov ako napr. v práci [7]. Avšak v súčasnosti tieto modely neponúkajú dostatočnú analýzu a preto sa do popredia dostávajú modely, ktoré vytvárajú hierarchiu tém. Medzi takéto modely patria napr. [8, 9].

4 Navrhované prístupy

V tejto kapitole predstavíme možné prístupy k modelovaniu témy pomocou FCA. Ako už bolo spomenuté FCA hľadá závislosti medzi vstupnými dátami a vytvára hierarchickú štruktúru. Avšak problémom pri aplikácii na úlohu modelovania témy je, že

FCA v základnej forme nedokáže odkrývať skryté vzťahy medzi vstupnými dátami, dokáže odkrývať len už známe závislosti.

Predpokladáme, že daný problém by bolo možné odstrániť 2 možnými spôsobmi. Prvý spôsob predstavuje použitie FCA v spojení s externou metódou, ktorá odkrýva skryté štruktúry (napr. latentné sémantické metódy) a FCA bude slúžiť len na vybudovanie hierarchie. Avšak ak by mal byť tento prístup využitý na spracovanie prúdov dát musí byť použitá inkrementálna externá metóda.

Druhým možným prístupom je modifikácia FCA tak aby dokázala skryté štruktúry medzi vstupnými dátami nájsť napr. aplikáciou pravdepodobnostných metód a kombináciou s inými algoritmami strojového učenia.

Prvý prístup bol experimentálne overený na vzorke 1000 príspevkov zo sociálnej siete Twitter pojednávajúcich o 4 hlavných témach. Ako externá metóda v spojení s FCA bola použitá metóda SVD (rozklad na singulárne hodnoty) [10]. Tento prístup bol porovnávaný s klasickými prístupmi modelovania témy (LDA) a zhľukovania (K-means). Kvalita metód bola porovnávaná na základe čistoty konceptov (purity) a počtu vygenerovaných konceptov. Ako je možné vidieť z Tab 1 nami navrhovaná metóda ma porovnateľné hodnoty čistoty ako klasické prístupy. Veľký počet konceptov pri navrhovanej metóde je spôsobený vytvorenou hierarchickou štruktúrou. Viac o tomto prístupe bude možné nájsť v práci [11].

Tab. 1 Porovnanie štandardných metód s nami prezentovaným prístupom (FCA-SVD) s rozličnými nastaveniami odpadu (odstránené koncepty pokrývajúce menej % objektov ako stanovený prah) a K (najlepších K singulárnych hodnôt z SVD)

| Metóda | Odpad (v %) | K (SVD) | Čistota | Počet konceptov |
|---------|-------------|---------|---------|-----------------|
| LDA | - | - | 0,699 | 4 |
| K-means | - | - | 0,766 | 4 |
| FCA-SVD | 5 | 4 | 0,73 | 91 |
| FCA-SVD | 5 | 8 | 0,69 | 750 |
| FCA-SVD | 5 | 20 | 0,55 | 2942 |
| FCA-SVD | 10 | 4 | 0,72 | 68 |
| FCA-SVD | 10 | 8 | 0,67 | 455 |
| FCA-SVD | 10 | 20 | 0,49 | 784 |
| FCA-SVD | 0 | 4 | 0,74 | 185 |
| FCA-SVD | 0 | 8 | 0,72 | 1669 |
| FCA-SVD | 0 | 20 | 0,64 | 25383 |

5 Záver

V článku boli prezentované možné prístupy k hierarchickému modelovaniu témy s využitím formálnej konceptovej analýzy, kde jeden z prístupov bol aj experimentálne overený a dosahoval porovnateľné výsledky ako štandardné nehierarchické prístupy. V budúcnosti by sme chceli rozšíriť experimenty na väčšie datasety a nájsť vhodnejšie metricky porovnávanie hierarchie tém ako čistota a počet zhľukov.

Pod'akovanie: Tento príspevok vznikol s podporou slovenského VEGA projektu č.1/0493/16 a slovenských KEGA projektov č.025TUKE-4/2015 a č.014TUKE-4/2015.

Literatúra

1. Belohlavek, R.: Introduction to formal concept analysis. Palacky University, Department of Computer Science, Olomouc, (2008).
2. Kuznetsov, S. O., Obiedkov, S. A.: Comparing performance of algorithms for generating concept lattices. *Journal of Experimental & Theoretical Artificial Intelligence*, 14(2-3), 189-216, (2002).
3. Butka, P.: Zovšeobecnený jednostranne fuzzy konceptový zväz a jeho ekvivalencia s konceptuálnym škálovaním. In *Znalosti*, Praha: Nakladatelství Oeconomica, ISBN 978- 80-245-2054-4, (2014).
4. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3, 694–703 (2003).
5. Petterson, J., Buntine, W., Narayanamurthy, S., Caetano, T., Smola, A.: Word Features for Latent Dirichlet Allocation. *Adv. Neural. Inform. Process. Syst.* 23, 1921–1929, (2010).
6. Zhai, K., Boyd-Graber, J.: Online Latent Dirichlet Allocation with Infine Vocabulary. In: *Proc. ICML 2013*, Atlanta, US, 561-569 (2013).
7. Li, X., Ouyang, J., Lu, Y.: Topic modeling for large-scale text data. *Front. Electr. Electron. Eng.* 16(6), 457–465, (2015).
8. Blei, D., Griffiths, T., Jordan, M.: The nested Chinese restaurant process and Bayesian non-parametric inference of topic hierarchies. *J. ACM* 57(2), article number 7, 1-30, (2010).
9. Hofmann, T.: The cluster-abbreviation model: Unsupervised learning of topic hierarchies from text data. In: *Proc. of IJCAI99*, Stockholm, Sweden, 682–687, (1999).
10. Rajaraman, A., Ullman, J.D.: Mining of massive datasets. Cambridge University Press, Cambridge, (2012).
11. Smatana, M., Butka, P.: Hierarchical Topic Modeling Based on the Combination of Formal Concept Analysis and Singular Value Decomposition, *Advances in Intelligent Systems and Computing, Multimedia and Network Information Systems* (v tlači), (2016).

Annotation:

Hierarchical Topic Modeling on Streams of Social Networks Data Based on the Formal Concept Analysis

One of the possible ways to analyse unstructured texts is topic modelling, which is trying to uncover hidden thematic structures in these texts. Topic modelling can be particularly useful in the context of social networks, which can be used for analysis at the time of crisis situations etc. Some of the extensions in topic modelling are related to hierarchical modification of conventional approaches, which offer deeper analysis than classic topic modelling. The main aim of this paper is to describe a new ways of hierarchical topic modelling based on formal concept analysis. It also provide experimental evaluation of one proposed method.

Cluster Based Symbolization

Milan Spišiak, Ján Kollár

Faculty of Electrical Engineering and Informatics
Technical University of Košice
Letná 9, 042 00 Košice, Slovak Republic

{ milan.spisiak, jan.kollar }@tuke.sk

Abstract. Complex object recognition applies in many sectors. However, design of these methods is difficult. Because of this there are many approaches. One of these is cluster based symbolization. Cluster based symbolization shows interesting results in this area. The approach is able to recognize events like human motions or gestures. From specific point of view the approach is similar to human neural networks. There are bigger and smaller clusters that can be connected with other clusters via references. The clusters have varying attributes like average cluster size or symbol dispersion in cluster. In this paper we review these attributes. For this goal we perform experiments using image objects. These objects represent letters of informal alphabets.

Contribution type: PhD Symposium

Keywords: Complex object recognition, Cluster based symbolization, Symbolization

1 Introduction

Symbolization process stands for very importing role in the process of advance object recognition [2,6,7]. The process allows us to encapsulating a raw image data to abstract structures. Complex events or objects are easier analyze by abstract structures than the raw data. Therefore, we decide to focus on symbolization methods allowing the complex recognition. To data store we use a cluster approach. The cluster approaches may potentially allow us recognize unlearned data. These approaches were used for a complex recognition (Takano [6], Zhou [8]) and they had interesting results in this area. However, our suggested method uses object vectorization instead of the hidden Markov model to get symbols. The change of symbolization method could change properties of cluster system. Therefore, we mainly focus on cluster properties in the tests.

2 Implementation

We divided implementation process to several parts:

1. The first part represents image processing. Here we try to reduce image noise by using image denoising method described in [1,5]. The method has input attributes allowing us to control the level of image denoising. The attributes affect the speed of method execution. Therefore, it is necessary to choose right ratio between execution speed and denoising level. There is no method with 100% noise reduction. In our case residual noise is located around object edges. This residual noise is reduced in following steps.
2. The second part is image threshold. We use test images with well-defined background in the experiment, because of this we decided to use the method based on one threshold level [4]. The main problem is to find out ideal value of threshold. We have to calculate with the residual noise. Because our approach uses object vectorization process [3] for an object description, we may not be very precise in the defining of threshold value - the object vectorization process is designed to adapt a low level of noise around object edges. We use the threshold value 64 for each monochromatic color channel. Values higher than the threshold are filtered out.
3. The third part represents symbolization process. For this goal we use the method described in [3]. The method is based on the object vectorization. The method allows us to describe outside and inside shapes. The method by vectorization of shape edges gets a string. As we wrote in the second part the method is able to reduce a noise around edges. A level of noise adaptation can be changed as required by one input method attribute.
4. Final part creates system based on clusters. Previous step gives us an object representation as a string. And in this part we want to store these strings to clusters. For this aim we need a storing system. Based on Takano's [6] and Zhou's [8] works, we design system using measurement differences between strings. The process of string insertion to cluster needs a decision condition. The decision condition defines whether a string will be inserted to cluster, or not. If we use comparison based on comparing all characters in strings, the system will be slow when it is bigger. Therefore, we decided to calculate a hash value for each string. The decision condition uses this hash values for comparison instead of comparison of all string characters. If no cluster satisfying the decision condition exists, a new cluster is created. To control cluster size, we define a new attribute – deviation value. This value stands for maximal difference between a string hash and a cluster hash. The cluster hash value equals to the string hash value of the first inserted string. The hash function is:

$$f(S) = \sum_{i=1}^n ((i+1)s(i)) \quad (1)$$

Here S stands for a string on input, n is the number of symbols in this string, $s(i)$ represents i^{th} symbol in this string, and $(i+1)$ is an increment providing a better cluster diversity. Excluding increment $(i+1)$ from this formula, the number of clusters decreases and the number of strings in clusters increases.

Two clusters are stored in the same cluster if absolute difference value of their hash values is lower or equal to the deviation value. However, if cluster contains two or more strings, it is calculated the average value of all hash string values in cluster. And the average value is used for comparison. The second method provides no redundancy in clusters. The method uses following formula:

$$f(S1, S2) = \sum_{i=1}^{\max(m1, m2)-1} abs(s1(i) - s2(i)) \quad (2)$$

Here function max returns maximum of lengths $m1$ and $m2$, abs returns the absolute value of its argument, and functions $s1(i)$ and $s2(i)$, are defined in range $i=0 \dots max(m_1, m_2)-1$ as follows:

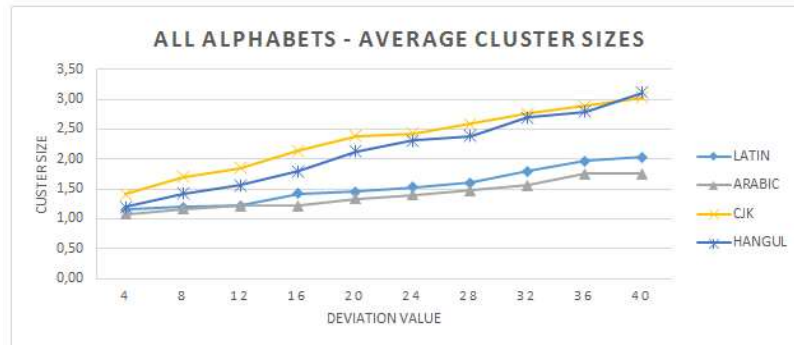
$$\begin{aligned} s1(i) &= s1(i), \text{ if } m1 \geq m2 \parallel m1 < m2 \ \&\& \ i < m1 \\ s1(i) &= 0, \text{ if } m1 < m2 \ \&\& \ m1 \leq i < m2-1 \\ s2(i) &= s2(i), \text{ if } m2 \geq m1 \parallel m2 < m1 \ \&\& \ i < m2 \\ s2(i) &= 0, \text{ if } m2 < m1 \ \&\& \ m2 \leq i < m1-1 \end{aligned}$$

3 Experimental result

Now we want to present results of our method on test data. In tests we mainly focus on these specific attributes of clusters:

1. Average cluster size – this attribute represents average number of strings in cluster.
2. Average strings dispersion – this attribute shows average dispersion in clusters. The ability to differentiate two strings from each other's depends on the difference between these strings. Higher difference means better recognition ability. Dispersion value shows this difference between strings in clusters.
3. Number of clusters - this attribute represents actual number of clusters in system.

We can affect tested attributes through the deviation value. We continually changed the deviation value and observed changes in the tested attributes. We mainly focused on type of relations between the deviation value and the tested attributes. Based on the type of relations we can estimate a future system behavior. In general, two system properties are problematic: unambiguous recognizing of symbols over time and a system cost. Measured values create linear relations. That's mean the system cost will grow linearly with growing data. And the average strings dispersion is linearly reduced to a point that it is not acceptable from the point of unambiguous recognizing of symbols. Through linearly relations the point of potential ambiguity is good estimable over time. Experimental results are shown on Obr. 1.



Obr. 1 All alphabets –Average cluster sizes

4 Conclusion

In this paper we present our approach for symbolization based on clusters. We performed several experimental tests in order to find out the system properties. Results are presented in section 3. Results show potential in the object recognition. However, this area needs further research.

Acknowledgment: This work was supported by project KEGA 031TUKE-4/2016 "Integrating software processes into the teaching of programming".

Bibliography

1. Buades A, Coll B, Morel JM. A non-local algorithm for image denoising. In *Computer Vision and Pattern Recognition*, (2005), pp. 60-65.
2. Kollár J. Automated language symbolization and conceptualization in human – computer communication, (2015)
3. Kollár J, Spišiak M. Direction vector grammar. In *Scientific Conference on Informatics*, (2015), pp. 151-155.
4. Rosin PL, Ioannidis E. Evaluation of global image thresholding for detection. *Pattern Recognition Letters*, (2003), pp 2345-2356.
5. Sonka M, Hlavac V, Boyle R. *Image processing, analysis, and machine vision*. (2014).
6. Takano W, Yamane K, Nakamura Y. Capture database through symbolization, recognition and generation of motion patterns. In *Robotics and Automation*, (2007), pp. 3092-3097.
7. Yongjun W, Guojie J, Suyvan H, Yingdong C. Research on 3D symbolic representation of geographical information, (2010), pp 1-4
8. Zhou X, Zhou X, Bouguettaya A, Taylor JA. A subspace symbolization approach to content-based video search. In *Data Engineering, ICDE'09*, (2009), pp. 1191-1194.

Predikcia Parkinsonovej choroby pomocou signálov reči použitím metód dolovania v dátach

Michal Vadovský, Ján Paralič

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky
Technická univerzita v Košiciach
Letná 9/B, 042 00 Košice, Slovenská republika

michal.vadovsky@tuke.sk, jan.paralic@tuke.sk

Abstrakt. Pri určovaní symptómov a predikovaní vybraných chorôb v medicíne sa často používajú zdravotné výsledky pacientov, ktoré sú získané z rôznych testov. Pri ľuďoch trpiacich Parkinsonovou chorobou existuje viacero príznakov, avšak typickým symptómom je problém s artikuláciou a rečou (dysfónia). Práve preto sme sa v tomto článku zamerali na klasifikáciu pacientov podľa ich rečových signálov použitím metód dolovania v dátach (Naivný Bayesovský klasifikátor a rozhodovacie stromy – algoritmy C4.5, C5.0 a CART). Dátová množina s ktorou sme pracovali sa skladá z hlasových meraní 31 osôb, pričom každá z nich má v dátach zastúpenie približne 6 záznamami. Najprv sme rozdelili dáta na tréningovú a testovaciu množinu a po vytvorení modelov sme vypočítali ich presnosť z hodnôt v kontingenčnej tabuľke. Okrem toho sme tiež pomocou hypotéz sledovali závislosť cieľového atribútu v binárnom tvare ku ostatným atribútom.

Typ príspevku: Doktorandské sympóziu

Kľúčové slová: Parkinsonova choroba, reč, dolovanie v dátach, klasifikácia

1 Úvod

Parkinsonova choroba [1] je druhým najčastejším neurodegeneratívnym ochorením hneď po Alzheimerovej chorobe. Vo vyspelých svetových krajinách sa vyskytuje približne u 0,3% populácie, pričom rastúcim vekom sa toto percento postupne zvyšuje. U ľudí starších ako 60 rokov už hovoríme približne o 1% a po veku 80 rokov dokonca o 4% ľudí z celkovej populácie. Z pohľadu pohlavia sa častejšie vyskytuje u mužov v pomere 3:1, čo môže súvisieť hlavne s ochrannými účinkami estrogénu u žien.

Príznaky tejto choroby sa líšia u každého jednotlivca individuálne. Jednotlivé symptómy sa u niektorých prejavujú pomaly, u iných zas rýchlejšie. Typickými prvotnými príznakmi môžu byť napríklad trasenie rúk, paží, nôh, ale aj spomalenie pohybu, stuhnutie svalstva a problémy s rečou [2]. V súčasnej dobe však neexistuje vhodná metóda liečby, ktorá by dokázala pacientov trpiacich touto chorobou úplne vyliečiť. Aspoň

z časti im pomáhajú lieky, ktoré nahrádzajú chýbajúci dopamín, vďaka ktorému sú pacienti udržiavaní v dobrej kondícii.

V našom výskume sme sa zamerali na diagnostiku Parkinsonovej choroby pomocou transformovaných dát zo zvukových záznamov do atribútov vyjadrujúcich signál reči. Táto oblasť je čoraz viac využívaná, pretože systémy založené na spracovanie signálov reči sú menej nákladné a jednoduchšie na použitie [3]. Tento prístup môže napomáhať včasnej diagnóze choroby. Hlavným cieľom práce bolo zistiť, akú presnosť dosiahnu vytvorené modely z dát rečových signálov a následné zhodnotenie ich reálnej použiteľnosti v praxi pre klasifikáciu pacientov.

2 Podobné práce

Parkinsonova choroba sa vyskytuje u ľudí veľmi často a napriek tomu stále na ňu neexistuje žiaden liek. Preto sa množstvo výskumníkov zameriava práve na túto oblasť. Napríklad v práci [4] zozbierali od 40 ľudí spolu 1040 nahrávok (26 vzoriek jedného človeka), z ktorých 20 ľudí trpelo touto chorobou. Tieto zvukové nahrávky obsahovali spracovaný signál reči z vyslovovania vytrvalých samohlások (a, o, u), čísel od 1 do 10, krátky viet a určitých slov. Okrem týchto atribútov obsahovali dáta pre každý záznam aj hodnotu UDPRS, čo je vlastne unifikovaná škála pre hodnotenie Parkinsonovej choroby určená odbornými lekármi. Ich cieľom bolo zistiť, aký typ hlasového záznamu (vytrvalé samohlásky, čísla, vety) dokázu lepšie predikovať túto chorobu alebo či transformovanie viacerých záznamov pacienta do určitých súhrnných a rozptylových metrick dokáže poskytnúť lepšie výsledky modelov. Záverom tohto výskumu bolo to, že najvyššiu presnosť dosiahla vytrvalá samohláska „o“ (72,5%) a slovo „four“ (75%). Okrem toho taktiež zistili, že prezentovanie viacerých záznamov jedného pacienta v súhrnných a rozptylových metrikách (medián, štandardná odchýlka, medzikvartilový rozsah a priemerná absolútna odchýlka) zlepšil zovšeobecnenie prediktívneho modelu. Prezentovaním pacienta pomocou priemeru a štandardnej odchýlky získali 82,14% presnosť modelu.

Sledovanie progresu ochorenia pomocou hodnoty UPDRS sa venovali autori v práci Tsanas s kol. [5]. Z dostupných dát, ktoré obsahovali rovnako signály reči sa pokúšali pomocou rôznych typov regresie predikovať hodnotu UPDRS. Pri predikovaní numerického atribútu *Motor-UPDRS* pomocou metódy CART dosiahli na testovacích dátach najmenšiu absolútnu chybu (MAE) s hodnotou 5,8.

3 Popis dát a modelovanie

Dáta, s ktorými sme pracovali vytvoril Max Little z Univerzity v Oxforde v spolupráci s Národným centrom pre hlas a reč sídlacim v Denvery v štáte Colorado a sú voľne dostupné na internete v databáze UCI Machine Learning Repository [6]. Celá množina dát obsahovala záznamy 31 pacientov, pričom 23 z nich trpelo Parkinsonovou chorobou [7]. Spolu bolo k dispozícii 165 záznamov (riadkov), pretože každý pacient mal v dá-

tach viacero záznamov, ktoré boli brané nezávisle od seba. Cieľový atribút bol s názvom **status** a obsahoval binárne hodnoty 1/0, pričom hodnota 1 znamená diagnózu Parkinsonovej choroby. Dáta obsahovali tieto atribúty: Meno pacienta a číslo nahrávky (*Name*), priemerná základná vokálna frekvencia (*MDVP:F0(Hz)*), maximálna základná vokálna frekvencia (*MDVP:Fhi(Hz)*), minimálna základná vokálna frekvencia (*MDVP:Flo(Hz)*), merania variability v základnej frekvencii (*MDVP:Jitter(%)*, *MDVP:Jitter(Abs)*, *MDVP:RAP*, *MDVP:PPQ*, *Jitter:DDP*), merania variability v amplitúde (*MDVP:Shimmer*, *MDVP:Shimmer(dB)*, *Shimmer:APQ3*, *Shimmer:APQ5*, *MDVP:APQ*, *Shimmer:DDA*), merania pomeru hluku a tónových zložiek v hlase (*NHR*, *HNR*), zdravotný stav pacienta (*Status*), nelineárne dynamické merania komplexnosti (*RPDE*, *D3*), signál fraktálovo-škálovateľného exponentu (*DFA*), nelineárne merania variability zákl. frekvencie (*spread1*, *spread2*, *PPE*). Načítanie dát, úprava dát a všetky experimenty (sledovanie závislosti medzi atribútmi, vytváranie modelov) boli vytvorené v prostredí RStudio pomocou programovacieho jazyka R.

Pre sledovanie závislosti medzi cieľovým atribútom (*status*) a všetkými ostatnými atribútmi, ktoré boli numerické, sme použili metódu testovanie hypotéz. V zostavených hypotézach sa sledovala podobnosť medzi priemerami atribútov rozdelených podľa cieľového binárneho atribútu. Vytvorili sme si nultú a alternatívnu hypotézu:

- H_0 : Priemer pacientov dvoch množín je rovnaký (nezávislosť atribútov)
- H_A : Medzi priemerami pacientov existuje rozdiel (závislosť atribútov)

Na testovanie týchto dvoch hypotéz sme použili Welchov dvojvýberový t-test, pri ktorom sledujeme hlavne hodnotu p (p -value). Čím nižšia je táto hodnota, tým je väčšia pravdepodobnosť závislosti cieľového atribútu so zvoleným numerickým atribútom. Vo všetkých prípadoch nám vyšla veľmi nízka p hodnota, čiže môžeme povedať, že existujú závislosti medzi cieľovým a ostatnými atribútmi. Napríklad medzi *Status* a *MDVP:Fhi* vyšla p -hodnota = 0,028, čo znamená, že na $(1-p)*100$ percent môžeme zamietnuť H_0 a potvrdiť H_A – v tomto prípade max. s 97,2% dôverou zamietame H_0 .

Vytvorenie modelov pre predikciu Parkinsonovej choroby sme robili pomocou metódy rozhodovacích stromov (algoritmus C4.5, C5.0, CART) a Naivného Bayesovského klasifikátora. Pred samotným modelovaním sme dáta rozdelili najprv na tréningovú a testovaciu množinu v pomeroch 70:30 a 80:20. Hodnoty získané modelmi boli porovnávané s hodnotami cieľového atribútu v testovacej množine, vďaka čomu sme vypočítali presnosť modelov pomocou vzorca $P = TP + FN / (TP + FP + TN + FN)$. Tento vzorec vyjadruje pomer správne klasifikovaných záznamov ku všetkým záznamom v dátach. Dosiahnuté výsledky uložené v Tab. 1 sme získali na dátach z testovacej množiny. Môžeme si všimnúť, že najlepšie výsledky dosiahol algoritmus C4.5 s úspešnosťou 91,43% pri rozdelení dát v pomere 80/20. Je potrebné si uvedomiť, že pri rozdelení dát v pomere 80/20 sme mali k dispozícii viac dát na tréning vytváraného modelu, ktorý tak mohol zachytiť viac vzorov v dátach a lepšie klasifikovať dáta z testovacej množiny. Najhoršie výsledky mal Naivný Bayesovský klasifikátor pri oboch rozdeleniach dát. V prípade použitia metódy zhukovania sme dosiahli úspešnosť iba 73,85%. Táto metóda rozdelila záznamy podľa podobnosti hodnôt ich atribútov do dvoch zhukov (tried), kde tieto triedy boli porovnané so statusom pacientov (1 – pacienti trpiaci Parkinsonovou chorobou, 0 – zdraví pacienti).

Tab 9. Dosiahnuté výsledky použitých metód

| Rozdelenie dát | Rozhodovacie stromy | | | Naivný Bayesovský klasifikátor |
|----------------|---------------------|--------|--------|--------------------------------|
| | C4.5 | C5.0 | CART | |
| 70/30 | 90,90% | 81,81% | 86,36% | 62,12% |
| 80/20 | 91,43% | 88,57% | 82,86% | 77,14% |

4 Záver a budúca práca

V tomto článku sme popísali prvé experimenty a modely pre predikciu Parkinsonovej choroby na dátach získaných zo záznamov reči pacientov. Podľa výsledkov v Tab 1. si môžeme všimnúť, že metóda rozhodovacích stromov a jej algoritmus C4.5 dosiahol veľmi dobré výsledky, dokonca pri oboch rozdeleniach dát má presnosť nad 90%. V publikácii [8] bola na rovnakých dátach dosiahnutá najvyššia presnosť 76% pomocou metódy Support Vector Machine. Môžeme teda povedať, že klasifikovať pacientov podľa transformovaných ukazovateľov (atribútov) ich reči je možné v celku úspešne. Testovanie hypotéz nám taktiež dokázalo, že všetky získané atribúty zo záznamov reči boli dôležité pre výslednú klasifikáciu pacienta.

V budúcej práci by sme sa chceli zamerať na tento typ choroby a riešiť ďalšie výzvy v tejto oblasti. Dôležité je nielen správne určiť, či pacient trpí Parkinsonovou chorobou, ale aj v akom štádiu sa nachádza. Najťažšie bude podľa všetkého správne určiť počiatkové štádium tejto choroby, kde predpokladáme, že sa jednotlivé ukazovatele reči nebudú až tak líšiť od zdravých ľudí. Určenie štádia choroby chceme riešiť aj výpočtom ukazovateľa UPDRS, napríklad pomocou rôznych metód regresie keďže sa jedná o numerický atribút. Na internete sú k dátam v tabuľke dostupné aj nahrávky, z ktorých boli získané jednotlivé atribúty. Tieto atribúty boli z reči pacientov transformované pomocou softvéru *Praat Acoustic Analysis*, ktorý je voľne dostupný na internete. Taktiež by sme chceli skúsiť aj iné softvéry, resp. získať aj iné ukazovatele, pomocou ktorých by sme dokázali vytvoriť lepšie predikčné modely s vyššou presnosťou.

Podakovanie. Táto publikácia vznikla vďaka podpore Vedeckej grantovej agentúry MŠVVaŠ SR a SAV projekt č. 1/0493/16.

Literatúra

1. Dexter, D.T., Jenner, P.: Parkinson disease: from pathology to molecular disease mechanisms. *Free Radical Biology and Medicine*, (2013), vol. 62, pp. 132-144.
2. Cnockaert, L., et al.: Low-frequency vocal modulations in vowels produced by Parkinsonian subjects. *Speech Communication*, (2008), vol. 50, no. 4, pp. 288-300.
3. Little, M.A., et al.: Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, (2009), vol. 56, no. 4, pp. 1015-1022.
4. Sakar, B. E., et al.: Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*, (2013), vol. 17, no. 4, pp. 828-834.

5. Tsanas, A., et al.: Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*, (2010), vol. 57, no. 4, pp. 884-893.
6. UCI Machine Learning repository: Center for Machine Learning and Intelligent Systems – Parkinsons Data Set. Available at: <https://archive.ics.uci.edu/ml/datasets/Parkinsons>.
7. Little, M.A., et al.: Exploiting Nonlinear Recurrence and Fractal Scalling Properties for Voice Disorder Detection. In: *Biomedical Engineering Online*, (2007), vol. 6, no. 23.
8. Geeta, Y., et al.: Predication of Parkinson's disease using data mining methods: A comparative analysis of tree, statistical and support vector machine classifiers. In: *Computing and Communication Systems (NCCCS), 2012 National Conference on, IEEE*, (2012), pp. 1-8.

Annotation:

Parkinson's disease Symptoms Prediction using the speech signals in the data mining methods.

Health records of patients sourced from various testing methods are frequently used in medical field for symptoms determination as well as for selected diseases probability prediction. There are numerous symptoms among the population suffering from Parkinson's disease, however dysphonia – changes in speech and articulation – is the most significant precursor. This is the reason why the article is focused on patients classification based on their speech signals using the data mining methods (Naive Bayes classifier and decision trees – algorithms C4.5, C5.0 and CART). The Dataset applied in the article consists of 31 individuals' voice measurings, with each of the individuals being represented by circa 6 records within the set. The dataset was primarily split into the training and testing sets, followed by the models implementation. The accuracy of the values obtained employing the models was calculated using the contingency table. In addition, the binary format target attribute dependency upon other attributes was examined using hypotheses.